

**Emotions, Cognition, and Moral Philosophy**

Thesis

presented to the Faculty of Arts

of

the University of Zurich

for the degree of Doctor of Philosophy

by

Giuseppe Ugazio

from Italy

Accepted in the autumn semester 2012 on the  
recommendation of Prof. Dr. Peter Schaber and Prof. Dr. Claus Lamm

(2012)

# Giuseppe Ugazio

May 2012

<b>Acknowledgments .....</b>	<b>5</b>
<b>Abstract .....</b>	<b>6</b>
<b>1 General Introduction.....</b>	<b>8</b>
<b>Origins of Moral Psychology.....</b>	<b>9</b>
<b>1.2 The Social Intuitionist Model.....</b>	<b>11</b>
<b>1.3 The Emotion Constitution Model.....</b>	<b>15</b>
<b>1.4 The Universal Moral Grammar Theory .....</b>	<b>21</b>
<b>1.5 Philosophical Origins of the Trolley Dilemma .....</b>	<b>27</b>
<b>1.5.1 The Dual Process Theory of Moral Judgment .....</b>	<b>31</b>
<b>1.6 Rationale .....</b>	<b>36</b>
<b>2 The Role of Emotions for Moral Judgments Depends on the Type of Emotion and Moral Scenario. ....</b>	<b>38</b>
<b>2.1 Abstract .....</b>	<b>38</b>
<b>2.2 Introduction .....</b>	<b>38</b>
<b>2.3 Methods.....</b>	<b>45</b>
<b>2.4 Results .....</b>	<b>51</b>

<b>2.5 Discussion.....</b>	<b>57</b>
<b>3 The Causal Role of the LPFC in Social Norm Compliance. ....</b>	<b>63</b>
<b>3.1 Abstract.....</b>	<b>63</b>
<b>3.2 Introduction.....</b>	<b>63</b>
<b>3.3 Methods.....</b>	<b>67</b>
<b>3.4 Results.....</b>	<b>72</b>
<b>3.5 Discussion.....</b>	<b>77</b>
<b>4 Pragmatic implications of empirically studying moral decision-making ....</b>	<b>81</b>
<b>4.1 Abstract ....</b>	<b>81</b>
<b>4.2 Introduction.....</b>	<b>81</b>
<b>4.3 How ought we to act? ....</b>	<b>82</b>
<b>4.4 How do we act? ....</b>	<b>84</b>
<b>4.4.1 Moral Action.....</b>	<b>84</b>
<b>4.4.2 Moral Judgment.....</b>	<b>85</b>
<b>4.4.3 Neural Underpinnings.....</b>	<b>86</b>
<b>4.5 People do not behave in a way they ought to.....</b>	<b>90</b>
<b>4.5.1 Biased behaviour.....</b>	<b>91</b>
<b>4.5.2 Emotionally influenced behaviour.....</b>	<b>92</b>
<b>4.6 Why do we not behave in morally decent ways? ....</b>	<b>93</b>

<b>4.7 Improving moral behaviour.....</b>	<b>97</b>
<b>4.7.1 Nudging.....</b>	<b>97</b>
<b>4.7.2 Training.....</b>	<b>98</b>
<b>4.7.3 Education.....</b>	<b>99</b>
<b>4.7.4 Pharmacological enhancement.....</b>	<b>100</b>
<b>4.7.5 tDCS/TMS.....</b>	<b>101</b>
<b>4.8 Should we try to improve, and is it possible? .....</b>	<b>103</b>
<b>4.9 Conclusion.....</b>	<b>107</b>
<b>5 General Discussion.....</b>	<b>107</b>
<b>6 References.....</b>	<b>112</b>

## **Acknowledgments**

This work has been possible solely thanks to the support I received by the many supervisors who guided me in these three years of intense research work. Firstly, I thank Professor Claus Lamm who with (a lot of) patience shaped me into an experimental psychologist during the years we spent together. His commitment to scientific research and his wise suggestions have provided me with solid milestones for my future career.

I thank Professor Tania Singer, for introducing me to the scientific world and helping me thoroughly during my first two years of research work, and Professor Christian Ruff for guiding my first steps in neuroscientific and brain stimulation research in these last two years.

Thanks also go to Professor Ernst Fehr, who always believed in my capacities of becoming an empirical philosopher, for his advice and constant encouragements.

Further thanks go to Professors Peter Schaber and Ingolf Dalferth, for their help in fostering my philosophical ideas, and to Giorgia Silani, Karl Treiber, Adrian Etter, Nina Spiri, Claudia Paixao, Johanna Espin, Anthony Schlaepfer, Tamara Herz, Sally Gschwend and all the members of the “Foundations of Social Behavior” group, for their valuable technical support and very pleasant working experiences.

I gratefully acknowledge the financial support I received from the research priority program at the University of Zurich “Foundations of Social Behavior”.

Finally I also thank my wife Catalina and my parents for everything, especially for being always at my side in the moments of most need.

## Abstract

In this dissertation I propose a combined philosophical and scientific approach to morality with the intent of clarifying some of the most controversial issues in this field. The scientific contribution of my work will mostly focus on advancing our understanding of the role played by emotional (in study I) and cognitive mechanisms (in study II) in determining moral decisions. The philosophical component of this thesis will focus on discussing the implications of empirically studying morality, such as the two previously mentioned, for the development of more accurate philosophical moral theories. The introduction will introduce the existing descriptive theories and models of moral judgments, discussing their strengths and shortcomings. This thorough discussion allows to accurately design studies which may overcome the shortcomings found.

In study I, I demonstrate using behavioral measurements that emotions have an important role in moral judgments. More specifically, the influence they exert on moral judgments can be predicted when taking into account their motivational dimension. While emotions resulting in approach motivation increase the number of actions we judge as morally permissible, those resulting in withdrawal motivation reduce the number of actions we judge as morally permissible. Furthermore, study I also shows that the influence emotions exert on moral judgments depends on the type of moral decisions.

Study II assessed the role of the lateral pre-frontal cortex (LPFC) in moral behavior, using the method of transcranial direct current stimulation (tDCS). The LPFC is a brain region associated with cognitive control, and study II shows that it plays a *causal* role in motivating morally appropriate behavior. More specifically, the results show that LPFC mediates the sensitivity of individuals to the punishment threat associated with disobeying moral norms. Higher LPFC brain excitability achieved via anodal tDCS resulted in an increased compliance with the moral norm (fairness in this case), when a punishment threat is present; conversely, lowered brain excitability (cathodal tDCS) resulted in an increased norm violating behavior. Notably, this influence on behavior is specific for social contexts, i.e., when humans interact with each other, as a control condition without such a context showed no effects of LPFC stimulation.

The section that follows, a more traditional philosophical one, discusses the pragmatic implications of empirically studying morality (in studies such as the previous two) for moral philosophy. In particular, I will discuss how the empirical sciences could be used to inform moral thinkers from proposing theory founded on irrational or fallacious reasoning, as for instance being influenced by a cognitive bias. Further, an additional question addressed in this section concerns the morality of using our scientific knowledge to influence moral decisions, for instance promoting a more equal society influencing people's choices of how to distribute public money within the society.

In conclusion I discuss that a comprehensive model describing moral behavior, extending the existing ones, has to take into account the fresh evidence illustrated in studies I and II, i.e.: moral decisions are informed by several psychological processes, which have a variable involvement depending on the context in which the moral decisions are taken. The resulting model will allow for both emotions and cognitive mechanisms to play a role in moral judgment, suggesting that the involvement of one and/or the other strongly depends on the context in which moral decisions are taken.

## 1. General Introduction

Moral decisions are at the core of human societies. They regulate many crucial aspects of social interactions by specifying the range of permissible and forbidden actions for a given situation. Due to their relevance for human life, moral decisions have been the center of philosophical debates since the earliest stages of western philosophy, as for instance in Aristotle's *Nicomachean Ethics* (Roger, 2000). Since then, numerous theories of moral decision-making have been put forward. Many of these theories base normative statements of how one ought to decide on theoretical considerations of the nature of morality, and on descriptions of how moral decisions are taken (Hume, 1777/1960).

While it can be debated whether observations of moral behavior indeed allow normative inferences on how we should behave (Moore, 1903), having a clear understanding of how humans decide is of critical importance on a practical level. In fact, moral decisions with severe potential consequences pervade many aspects of human life. For instance, doctors routinely face such decision in the context of organ transplantations (e.g., if there are several candidate patients for this organ; Courtney & Maxwell, 2009). Other examples involve peacekeeping soldiers - who may have to decide whether to turn away refugees from an already full refugee camp - or coastguards who may have to decide whether to obey orders and refrain from rescuing drowning illegal immigrants (Krosch et al., forthcoming).

In the present work I combine a psychological approach to moral behavior with a philosophical discussion of the implications of empirical findings for the moral philosophical debate, with the aim of disclosing the necessity for moral psychology and philosophy to enter a dialogue in order to achieve these disciplines' ultimate goal: improve human social life.

In the first introductory section I give an overview of the philosophical and scientific literature describing morality and moral behavior, highlighting the different psychological processes involved in moral decision-making. This overview will put in evidence the importance of two main psychological features: *emotions* and *cognitive control*, and point out what needs to be clarified in order to significantly advance our understanding of moral behavior. Section two presents a behavioral experiment testing the role of emotions for moral judgments; section three examines the causal role of the lateral pre-frontal cortex in norm compliance; chapter four draws



a parallel between moral philosophy and psychology showing how these may complement each other in order to promote morally appropriate behavior; finally, chapter five concludes the dissertation with a general discussion and conclusions.

### **1.1. Origins of Moral Psychology**

In this section I take into account several descriptive theories of moral judgment, which mainly try to answer to the following core questions: What is the nature of morality? How do we form moral judgments? Which brain processes are involved in moral evaluation? Providing evidence to answer these issues motivated decades of research first in philosophy (Smith, 1759/2010; Hume, 1777/1960, 1785/1985), as well as in psychology (Prinz, 2006; Haidt, 2001; Greene et al., 2001; Kohlberg, 1984; Piaget, 1932), neuroscience (Mikhail, 2007; Greene et al., 2001), biology (Hauser, 2006), psychiatry (Cima et al., 2010; Huebner et al., 2008), and many other fields.

One of the first attempts to provide a complete description of moral judgments drawing on the methods of science is David Hume's *Treatise on Human Nature* (Hume, 1785/1985). The origins of modern moral psychology can be traced back to Hume's works, although the latter propose an empirical approach that can't be considered scientific given the contemporary scientific standards, i.e., an introspective analysis of one's own psychological state while elaborating moral judgments. Briefly, in the treatise on human nature, Hume delineated his moral descriptive theory stating that morality is of emotional nature: in order to evaluate the moral appropriateness of an event a person will ground her judgment on the "gut feeling" provoked by this event. If the gut feeling is a pleasant one, then the event is morally appropriate, otherwise if the gut feeling is unpleasant the event is morally inappropriate. In other words, paying attention to her own feelings a person can infer if something is morally appropriate or not.

In the 1930s, the Swiss psychologist Jean Piaget laid out his theory on child moral development. In this theory he described how children develop moral ideas in stages, shaping their moral judgments based mostly on their experience of the external world (Piaget, 1932). Drawing on this theory, Lawrence Kohlberg (Kohlberg 1984) developed a more detailed account of the stages a child would go through while developing his moral reasoning ability. Shortly, Kohlberg described six moral development stages divided in three levels: pre-conventional,

conventional and post-conventional. In the pre-conventional stages moral reasoning is merely determined by self-interest (e.g., how can I avoid punishment), while in the conventional stages moral reasoning is motivated by interpersonal interests (e.g., how can I contribute to maintain social order). Finally in the post-conventional stage, moral reasoning takes a more metaphysical stance being concerned with universal ethical principles. Both Kohlberg and Piaget state in their works that moral judgments stem from conscious reasoning.

In more recent years, empirical research on moral judgments has developed mainly around two research questions: a) do moral judgments stem from intuitions or from conscious reasoning, and b) which psychological processes are involved in moral intuitions (Cushman et al., 2010). Hereafter I focus mainly on the latter, discussing the different psychological theories put forward describing some of the processes taking part in moral judgments. More specifically, I will discuss the Social Intuitionist Model (Schnall et al., 2008a; Haidt, 2007, 2001; Wheatley & Haidt, 2005), The Emotion Constitutional Model (Prinz, 2006), the Universal Moral Grammar theory (Mikhail, 2007; Hauser, 2006), and the dual-process theory of moral judgment both in its original presentation (Greene et al, 2004, 2001) as well as in a re-elaborated form (Moll et al., 2008, 2007, 2005, 2002).

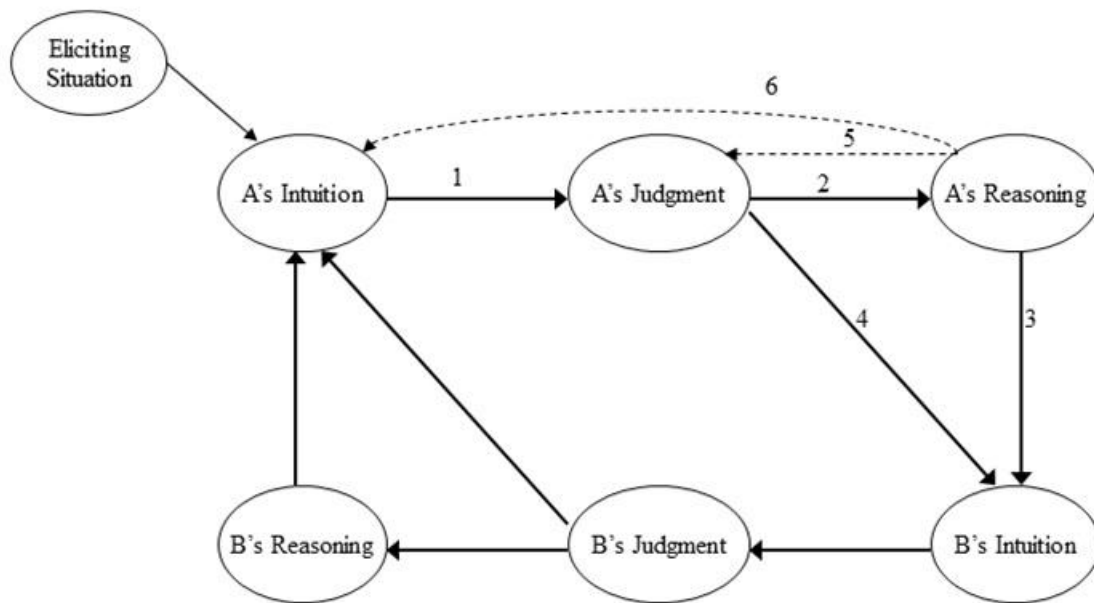
One of the main differences between all these models stands in the role emotions are believed to play in the determination of moral judgments. Briefly: for the first one, the Social Intuitionist Model, emotions are fully involved in the determination of all moral judgments; for the **ECM** emotions *constitute* moral judgments; for the **UMG** emotions play no role in determining the moral judgment, at most these can be a disturbing element but with no direct causal link to moral judgment; for the fourth and last one, emotions are involved in the determination of only some moral judgments.

Before proceeding to illustrate the different models it is however important to mention how the term *emotion* is used in this dissertation. This term, indeed, is being used in several different ways in the literature and might be the source of misunderstandings if not appropriately contextualized. Hence, in this dissertation the term emotion is used to denote affective reactions to an external stimuli which generate a feeling or mood in a person. Importantly, as used in this dissertation, emotions refer solely to basic emotions (Ekman, 1972), and do not take into

consideration complex emotions, which might be the product of combined cognitive and affective reactions (Lazarus, 1999).

## 1.2.The Social Intuitionist Model

The Social Intuitionist Model (see Figure 1) was proposed by Jonathan Haidt in 2001. This model, largely anchored to Hume’s moral theory (see above, 1785/1985 and 1777/1960), holds that moral judgments result from one’s “gut feelings”. Briefly, this model suggests that emotions, under the form of “gut feelings”, inform one about the moral appropriateness of an action, arising in a certain emotional state: a pleasurable one if the action is morally praiseworthy



**Figure 1** Schematic representation of The Social Intuitionist Model of moral judgment adapted from (10). The numbered links, drawn from Person A only, are 1) the intuitive judgment link, 2) the post-hoc reasoning link, 3) the reasoned persuasion link and 4) the social persuasion link. Two additional links are hypothesized to occur less frequently: 5) the reasoned judgment link, and 6) the private reflection link (see the original picture in Haidt, 2001, figure 2).

or an unpleasant emotional state if the action is morally wrong. Following this immediate reaction informing a person about the moral quality of an action, one expresses her moral judgment: if we are in a pleasant emotional state we will judge the action to be morally appropriate, otherwise to be morally inappropriate. Emotions are the most important element, being the direct source of information for moral judgments.

Conscious reasoning is excluded from the direct causes of moral judgments. According to this model, in fact, reasoning comes into play only after the moral judgment has been made, and has the main task of a) providing plausible justifications for the expressed judgment, and b) persuading others if necessary. Reason may still affect moral judgment though just in an indirect way by interacting with and exerting an influence on our emotional reactions (Haidt & Joseph, 2004). The influence on moral judgments is indirect in the sense that reason does not directly participate in the formation of moral judgments but solely in shaping emotions. For these reasons, the crucial hypotheses Haidt empirically tests in order to support with sound evidence his Social Intuitionist Model imply that a) moral judgments result mainly from intuitions, b) conscious reasoning has a secondary role, and c) moral intuitions are of emotional nature (Schnall et al., 2008a; Haidt, 2007; Wheatley & Haidt, 2005).

To test the first hypothesis, Haidt and colleagues (Schnall et al., 2008a, 2008b; Haidt, 2007; Wheatley & Haidt, 2005; Haidt & Joseph, 2004; Haidt, 2001) developed a series of moral vignettes describing violations of moral norms that are suggested to be strongly connected to feelings of disgust. For instance, one of the most representative scenarios describes two siblings who decide to have sexual intercourse:

Julie and Mark are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that? Was it ok for them to make love? (Haidt, 2001, p.1)

This vignette allowed testing the hypothesis that moral judgments are rather resulting from intuitions and not from conscious reasoning, as the vignette described above is constructed in such a way to exclude most, if not all, the possible reasons one may adduce to ground his moral judgment (e.g., the higher probability of having disabled offspring). Asking people to

express their moral judgment regarding the behavior of the two siblings and to provide with a brief explanation for their judgments, revealed that

Most of the people who hear the story immediately say that it was wrong for the siblings to make love, and they only subsequently begin searching for reasons. (Haidt, 2001, p.1)

In more details, the subjects interviewed by Haidt reported precisely what he hypothesized: the judgment on the action is informed by the “gut feelings” one is experiencing when exposed to such action. In this case for instance, according to Haidt, one feels disgust when reading the story of two siblings making love; furthermore from the statements provided by individuals taking part in the study (Haidt, 2001), it can be evinced that reasoning appears to be secondary; in fact reasons seem to be provided solely to justify the judgment, not to motivate it. For instance, some of the reasons reported by subjects are: it is wrong because they may have some psychological problem, or if she gets pregnant then the baby has a great chance of have genetic illnesses. As mentioned before, all these reasons one may adduce to motivate moral disapprove were a priori excluded by the phrasing of the vignette. Once participants were made aware of this fact, they acknowledged that the reasons provided were actually trying to justify the fact that they “just know it is wrong for the siblings to make love” (Haidt, 2001, p.1).

In this first study (Haidt, 2001) Haidt gathered evidence supporting his initial hypothesis that moral judgments are largely based on intuitions and only secondarily on conscious reasoning. According to the data collected so far, it seems that people first express their moral judgments based on their intuitions and only secondly find reasons to justify their judgment, and not vice-versa as others previously argued (e.g. Kohlberg, 1984; Piaget, 1932).

Following this initial evidence, Haidt and colleagues (Schnall et al., 2008a; Wheatley & Haidt, 2005) proceed to test the primacy of affect prediction, i.e. the hypothesis that moral intuitions are of emotional nature. In order to do so, they use the same kind of moral vignettes as the one described above. Other examples include: a congressman campaigning against corruption but accepting bribes; a person eating his own dead dog; a person counterfeiting his resume to enhance his chances of getting hired; allowing marriage between cousins. As previously mentioned, the authors hypothesize that the moral violations described in these vignettes are

strongly connected to feelings of disgust. Therefore, assuming that the primacy of affect prediction is correct, Haidt hypothesized that these scenarios would induce a feeling of disgust in the participants and that this feeling would influence the outcome of their judgment concerning whether the protagonists' action is morally permissible.

To obtain the evidence necessary to support this prediction, Haidt and colleagues used different techniques to induce disgust including disgusting smells (Schnall et al., 2008a) and hypnosis (Wheatley & Haidt, 2005). The evidence reported in these studies shows that the induction of disgust makes individuals' condemnation of moral violations such as incest harsher, compared to the judgments expressed by individuals in a neutral emotional state. Furthermore, the authors also report that the measured influence on moral judgments of an altered affect state is specific for disgust. In fact, using a similar procedure but this time inducing sadness in individuals, no effect on moral judgments was found (Schnall et al., 2008a).

Based on these empirical results Haidt draws some important conclusions on the nature of morality: first, that moral judgments result from intuitions and that reasoning has a marginal role in this matter; second, that human moral intuitions are driven by emotions; thirdly and most importantly, combining the two previous conclusions, Haidt argues that many of the moral norms pervading our societies directly result from a general psychological mechanisms evolved to process primary disgust in the human brain and body (Haidt, 2001).

To further strengthen the view that moral norms originated to prevent humans from engaging with behaviors that may lead to contamination, in an additional study Schnall and colleagues (2008b) hypothesized that inducing the feeling of purity in individuals, by increasing their physical cleanliness immediately before taking part in their study, would lead them to soften their moral condemnation of the actions described above (Schnall et al., 2008a; Wheatley & Haidt, 2005; and Haidt, 2001). The results obtained in this study confirmed that induction of a purity feeling leads to the opposite effect of disgust, i.e., it resulted in softened condemnation of moral violations.

In the paragraph that follows I introduce a moral philosophical theory proposed by Jesse Prinz (2007, 2006) which is largely grounded on the evidence discussed so far.

### 1.3. The Emotion Constitution Model

In several contributions (2007, 2006) to the moral philosophical debate mainly concerned with understanding the nature of morality (i.e. is morality a matter of emotions/passion or of beliefs/reason?), Prinz proposes the Emotion Constitution Model:

According to which emotions constitute moral judgments. [...] The moral judgment is not a further stage in processing following on the heels of the emotion, but is constituted by the emotion together with the action representation (Prinz, 2007, p.5).

More specifically, in his theory on moral judgment, Prinz elaborates the mentioned model by first defining moral concepts (such as right or wrong) as *sentiment*, which are the categorical basis of a disposition to experience different emotions (Prinz, 2007, pp.4-5). For instance:

The sentiment that constitutes the concept wrong disposes its possessor to feel emotions of disapprobation. If I judge that stealing is wrong, that judgment is constituted by the fact that I have a negative sentiment towards stealing a sentiment that disposes me to feel angry at those who steal and guilty if I myself steal. (Prinz, 2007, p.5)

Furthermore, according to this model, when a person needs to evaluate a given event, one will first analyze and categorize this event; secondly, if the categorized event corresponds to a moral sentiment, the former will generate an emotional reaction in the person evaluating the event. As a result, in the person's mind a moral judgment will arise which is constituted of a complex mental state combining on one hand the representation of the event and on the other hand the associated emotion.

As it can be noticed comparing the two models proposed so far, Prinz's Emotion Constitution Model can be seen as a radicalization of the Social Intuitionist Model: in fact if the latter holds that intuitions of emotional nature lead to moral judgments, the former states that moral judgments actually are constituted by the emotional intuitions. In other words, if for Haidt the ability to feel emotions is a *sufficient* condition for morality to exist, for Prinz such ability is

a sufficient and *necessary* condition for morality, and hence for moral judgments, to exist. Indeed, if an emotion constitutes moral judgments then in order to be able to make moral judgments the ability of feeling emotions is necessary to make moral judgments, or else one would have to allow for something other than emotion to result in a moral judgments. Due to this radicalization, the empirical predictions which need to be tested are now not only the ones entailed by the Social Intuitionist Model but at least also a crucial additional one, namely: moral judgments cannot exist without emotions.

In the following paragraph I will analyze the additional evidence Prinz refers to in order to empirically support his Emotion Constitution Model. To provide this evidence Prinz (2006) follows the following steps: first by showing that emotions are necessary to discriminate between moral and non-moral norms; second that one may consider something to be wrong solely because one has a negative feeling towards it; and finally that in the absence of emotions one is no longer able to make moral judgments.

Prinz begins his search for empirical support proposing a simple thought experiment, concerned with the difference between moral and non-moral rules. Consider the two following rules kids are frequently taught at schools: a conventional rule states that kids should raise their hands and wait for the teacher to give them the word before speaking; a moral rule states that pupils should not harm other kids. What makes the first a mere conventional rule and the second a moral rule? Prinz's argument is that when one thinks about transgressing the first rule, one may recognize the violation and think that it is bad to violate such a rule, but one does not have any negative feeling which makes one feel bad while violating such a norm. On the other hand, if a kid thinks about harming another one, he or she immediately has a negative emotional reaction (for instance of fear, guilt, or perhaps anxiety) informing him that the action he or she is thinking about (i.e., harming another kid) is wrong (2006).

Furthermore Prinz proposes another differentiation between the two types of rules. He argues (2006) that if we suppose a teacher at school tells the kids that they can freely talk whenever they want, therefore with no need to raise their hands and wait for having the word, most if not all of them will conform to the new norm, freely talking when they want to. On the contrary, if a teacher tells the kids that they may harm each other freely, very few of them would



actually do so. Prinz introduces this and other thought experiments as preliminary evidence appointing to the hypothesis that emotions are a crucial element discriminating between moral norms and non-moral norms, such as conventions. On one hand, conventional norms only exist as long as there is a sort of external (to an individual) enforcement. Once this enforcement is removed, the conventional norm will no longer exist. On the other hand, a moral norm will hold even once the external enforcement has been removed, as its existence depends mostly on the individuals' internal motivation. For Prinz, such motivation occurs via sentiments, which as stated before are of emotional nature (2007, 2006). If we consider the above example once the external authority, in this case the teacher, modifies the two norms, the kids would change their behavior with respect to the conventional norm, now freely talking when it pleases them, but not with respect to the moral norm, still abstaining from harming others.

The hypothesis that emotions are necessary to discriminate moral and non-moral norms found more solid scientific support beyond the thought experiments discussed so far. In fact, in two functional magnetic resonances imaging (fMRI) studies Moll and colleagues (2002) and Heekeren and colleagues (2005) provide evidence indicating that in human brains there are some moral-specific emotional neural networks. Very briefly, fMRI is a method which allows to indirectly observe which areas of the brain are active during a task by measuring the blood oxygenation level: the areas with higher oxygenation levels are the ones which are recruited by the brain to respond to stimuli, as the cells need more oxygen when they are in an active state with respect to when they are in a passive one.

Using this method Moll and colleagues (2002) measured in a study which brain areas are activated by morally relevant elements. In order to obtain these measurements the authors compared the brain activity registered in participants while they were observing pictures with several different contents: one type of pictures displayed unpleasantly emotionally charged content entailing morally salient elements (e.g., war scenes or physical assaults); a second type of pictures showed emotionally charged content, both pleasant (e.g. landscapes or people) and aversive (e.g., dangerous animals or body lesions), but with no morally salient elements; a third type included control pictures displaying emotionally neutral pictures, or scrambled images (e.g., landscapes or people).

The results obtained showed that pictures with unpleasant emotional content with moral and non-moral salience activated several brain areas associated with negative emotion processing (Damasio et al., 2000; Lane et al., 1999), as revealed by a conjunction analysis (an analysis identifying the brain areas commonly activated by the two types of pictures: emotionally charged with moral content; and emotionally charged without moral content) comparing these pictures against emotionally neutral images. In fact, viewing these pictures resulted in increased brain activity in the amygdala, the right thalamus, and the right insula/inferior frontal gyrus (Moll et al., 2002).

More importantly for Prinz's model, the analysis contrasting morally salient pictures against pictures with no moral content revealed increased brain activity elicited by the morally charged pictures in the orbitofrontal cortex (OFC), the superior temporal sulcus (STS), and the medial frontal gyrus (Moll et al., 2002). Together these regions have been proposed to be automatic detection of socially and emotionally salient events, or in Moll's words these regions are suggested to *be critical elements of a cortical–limbic network that enables humans to link emotional experience to moral appraisals* (Moll 2002, p. 2736). The results of this analysis thus seem to support the hypothesis that emotions are a fundamental element used to discriminate moral and non-moral norms, as there are brain mechanisms specific for morally salient elements within the emotional recognition network.

This hypothesis is moreover corroborated by the evidence Heekeren and colleagues (2005) obtained in a study measuring the changes in individuals' brain activity while they read texts either describing moral norms violations (A gives B a bloody nose) or containing a violation of a grammatical norm (A dresses a very bloody wound). This evidence revealed in fact that the moral violations alone activated a neural network of brain regions which included the posterior STS, , as well as the ventromedial prefrontal cortex (VMPFC), and the posterior cingulate cortex (PCC), this last being a region reported to be involved in processing the emotional relevance of stimuli (Cato et al. 2004; Maddock, 1999). In sum, the findings just introduced empirically partially corroborate the Emotion Constitution Model's prediction according to which emotions are necessary to discriminate between moral and non-moral norms. In the two studies introduced (Heekeren et al., 2005; Moll et al., 2002), in fact, it has been shown

that only scenarios including morally norms violations triggered emotional responses, while for instance factual scenarios do not. What the discussed data cannot disclose, however, is whether emotions are a) the only element necessary for discriminating between moral and non-moral norms, b) they are always necessary for this discrimination, or if they are necessary only for the specific moral decisions used to obtain this evidence, and c) relevant for the moral judgment is determined by the context, i.e. if the emotion is directed to something morally relevant, or by something else (see the next section for a more thorough discussion of the limits of this model and of the evidence used to support it). The next steps to validate the Emotion Constitutional Model are to discuss evidence showing that a) emotions are indeed able to induce a person to judge something devoid of any moral connotation to be morally inappropriate, and b) the absence of emotions results in morally inappropriate behavior (Prinz, 2006).

To support this step, Prinz relies on evidence provided for in the study by Wheatley and Haidt mentioned above (Wheatley & Haidt, 2005). In this study the authors tested if a disgust feeling induced via hypnosis would influence moral judgments, making these last more tough than in a control condition. Hence, individuals taking part in this experiment were induced via hypnosis to feel disgusted every time they would see the word “often”; then they were taken out of the hypnotic condition and presented with vignettes describing both morally condemnable characters but also morally praiseworthy ones. In one of these characters description proposed was introduced the word “often”, while in some other “often” was substituted by a synonym.

Individual’s behavior revealed that whenever the word “often” was present in the description of the character, their moral judgments about this person was always negative, even when the character was morally praiseworthy. In addition, morally condemnable characters when described using the word “often” were judged in a tougher way than when the word was substituted with a synonym. Therefore, these experimental findings suggest that emotions, in this case disgust, are sufficient to determine a moral judgment. In fact, the disgusted reaction towards a neutral word (i.e. often) was the only information about the character judged which was considered by subjects in order to express their moral judgments, even when they had good reasons to express praise towards the character described in the vignettes they read. With this evidence at hand Prinz considers the second step of his model validation concluded.

In order to provide evidence supporting the prediction that the absence of emotions results in morally inappropriate behavior, Prinz relies on findings proposed in the literature exploring the relationship between psychopathy and emotional capacities, particularly on the contributions of Blair and colleagues (2002, 1995). According to Prinz these contributions suggest *that emotions are developmentally necessary for acquiring the capacity to make moral judgments* (Prinz, 2006, P.32). More specifically, findings in the psychiatric literature suggest that psychopaths tend to engage more than healthy people in anti-social behavior and to have problems distinguishing between moral and conventional rules (Blair et al., 2002).

In other words this means that psychopaths do not have the capacity to distinguish between moral and non-moral violations: a psychopath can recognize that a certain action is violating a rule, but he is not able to distinguish if the action is violating a moral rule or a conventional one. Considering the thought examples discussed previously, the norm prohibiting talking without raising one's hand has no qualitative difference from the one prohibiting hurting another kid.

Such behavioral differences seem to result from a deficit in processing emotions observed in persons affected by psychopathy (Blair et al., 1995). Such a deficit often results in impairment to recognize emotional facial expressions or emotional intonations in speech. This general deficiency in processing emotions causes as well the inability of feeling moral emotions, hence explaining why psychopaths have difficulty understanding moral rules as well as the observed increased tendency of psychopaths, compared to healthy people, of engaging in anti-social behavior (Prinz, 2007; Blair et al., 2002). In sum these findings suggest that impairments in emotional capacities seem to result in an impaired ability of making moral judgments.

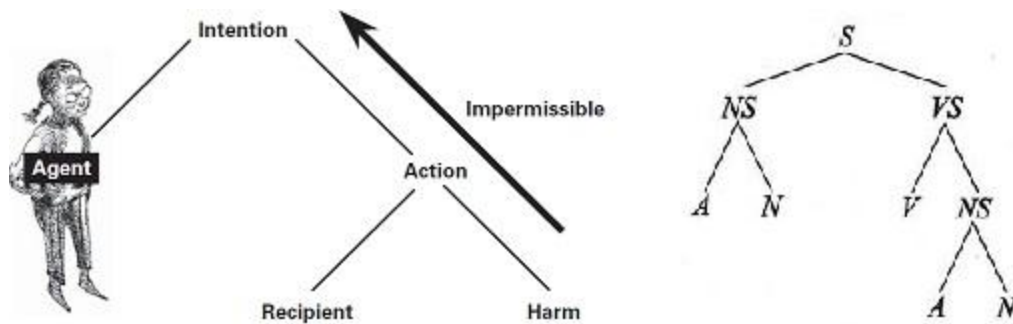
Summing up, in this section on the Emotional Constitution Model (Prinz, 2007, 2006) I provided an example of how empirical data can be used to suggest and support philosophical statements, in this specific case that morality is of emotional nature. The discussed data support the view that moral concepts and emotions might be strongly interconnected as it shows that emotions are sufficient and necessary for morality to exist. Sufficient as emotions allow to: a) discriminate between moral and non-moral rules (Heekeren et al., 2005; Moll et al. 2002), b) induce a person to morally condemn a certain situation based on his emotional intuition

(Wheatley & Haidt, 2005), and c) enhancing an emotional state (e.g. inducing disgust) directly affects the moral judgment associated with the given emotional state (Schnall et al., 2008a, 2008b; Wheatley & Haidt, 2005; Haidt & Joseph, 2004). Furthermore, emotions are necessary as the impairment of emotion processing results in impairments of moral judgment and behavior (Prinz, 2007; Blair et al., 2002, 1995). According to Prinz it is because of their emotional deficit that psychopaths treat moral norms as conventional ones.

However, it must be noted that evidence from several other moral psychological studies, discussed more specifically in the sections that follow, do not support all claims of the Emotional Constitutional Model's predictions. Briefly, in the moral psychological literature there is evidence that not all moral scenarios require an emotional response in order for an individual to elaborate a moral judgment (e.g. Greene et al., 2004, 2001). Being this the case, the requirement that emotions are necessary to generate moral judgments, in order to be held, must be eased to the following less demanding form: emotions are necessary only for certain types of morality. It is clear though that in its simplified version, the necessity condition has no longer the competence of supporting the statement that morality's nature is solely emotional, as if this was true all types of moral judgments and behaviors should be sensible to emotional variations. A theory considering moral judgments to be of purely emotional nature therefore seems therefore quite unlikely to be universally describing the nature of morality.

#### **1.4.The Universal Moral Grammar Theory**

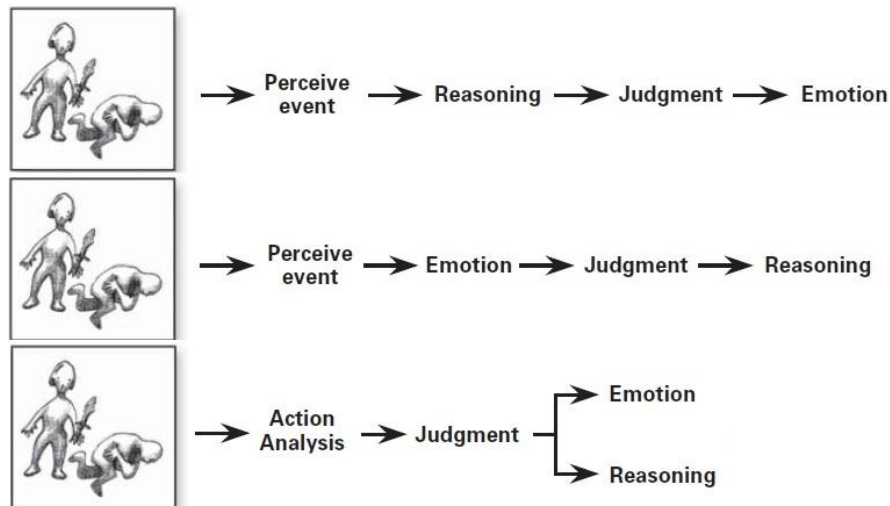
In this section I take into consideration a third alternative moral psychological theory named the Universal Moral Grammar. As pointed out by some supporters of this theory (Hauser et al., 2008), the roots of this theory can be easily identified a) in Chomsky's (2000, 1988, and 1986) linguistic theory known as the Universal Grammar as far as this moral theory's logical and b) in John Rawls' theory of moral judgment (1971) with respect to the theoretical framework determining the core hypotheses entailed in the Universal Moral Grammar model. In this theory, in agreement with the two previous models, the view is endorsed that moral judgments stem from intuitions (Mikhail, 2007; Hauser, 2006) and not from conscious deliberative reasoning (Kohlberg, 1984; Piaget, 1932) structure is concerned (see figure 2),



**Figure 2** The common logical system behind the Universal Moral Grammar on the left (Hauser et al., 2008, figure 3.5) and of the Universal Grammar on the right (Chomsky, 1986 figure 6).

Proponents of this view completely disagree with the former two theories with respect to the biological nature of moral intuitions. In fact the scholars backing the Universal Moral Grammar (Huebner et al., 2008; Mikhail, 2007; Hauser, 2008, 2006) deny any causal role for emotions in the processes giving rise to moral judgments. Instead, they claim that moral intuitions result from a morally specific psychological mechanism constituting the so-called *Moral Organ* (Hauser et al., 2008, 2006) of emotions and/or conscious reasoning. These moral mechanisms are responsible of implementing what the authors call the *action analysis*: i.e. the computations performed by the mentioned Moral Organ on a given situation required to generate a moral judgment in a person's head. More precisely, the action analysis mechanisms focuses on evaluating morally salient variables of a given situation, such as (see figure 2 on the left): who is the agent, what are his intentions (e.g. good/bad), what are his beliefs on the outcome generated by his action, what kind of action does the agent perform (e.g. help/harm), who is the receiver and what are the consequences (Hauser et al., 2008). Hence, evaluating all these elements the Moral Organ will produce a moral judgment, which may in turn generate emotional reactions or induce a reasoning process aimed at providing with justifications of the moral judgment just formulated (see figure 3). The Universal Grammar Model is supported by several studies providing empirical evidence corroborating some of the hypothetical requirements entailed by this model. In what follows I discuss the data presented by Hauser and colleagues to show that: a) manipulating two variables suggested to be evaluated in the moral organ's action analysis mechanism, intentions and outcomes, results in altered moral judgments (Young et al., 2010, 2006), and b) psychopaths' moral judgments do not differ from those expressed by healthy individuals (Cima et al., 2010). Furthermore in a review article (Huebner et al., 2008) the authors

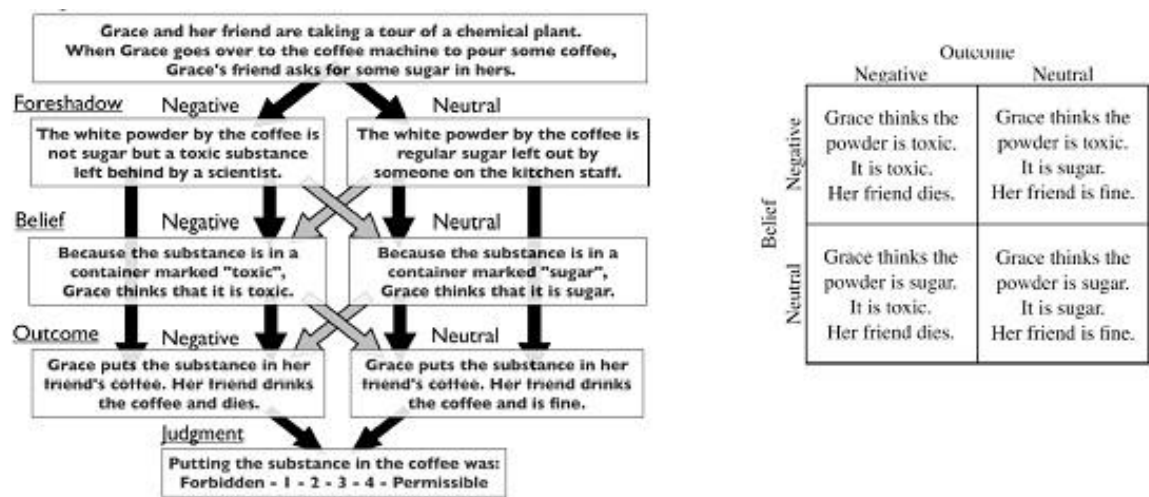
argue that emotions are not involved in the causal process of moral judgments formation proposing alternative explanations of the results discussed in the previous sections. The effects of manipulating intentionality and consequences on a person's moral judgments have been explored in a recent fMRI study (Young et al., 2006).



**Figure 3** A schematic representation of the three models discussed so far provided by Hauser, Young and Cushman (Hauser et al., 2008, figures 3.1, 3.2, and 3.3). On the top the model suggesting that moral judgments stem from reasoning (Greene et al., 2004, 2001); the middle one suggests moral judgments result from emotions (Prinz, 2006; Haidt, 2001); finally the bottom one claims that moral judgments result from an action analysis performed by the Moral Organ (Hauser, 2006).

In this study the authors introduced a new type of moral scenario describing a person knowingly (versus unknowingly) causing a negative outcome (e.g., someone's death), that is, believing that his actions will (versus will not) lead to the negative outcome. In more details, intentionality was manipulated modifying the knowledge possessed by the agents described in this type of scenario (see figure 4). Individuals taking part in this experiment were asked to read the scenarios and rate the permissibility of each moral scenario on a scale ranging from one to four. The evidence obtained in this study revealed that intentionality constituted the major source of information determining people's moral judgments as revealed by analyzing the scenarios describing attempted harm: namely when the intention described was negative but the outcome of harming someone was not achieved, resulting instead in a neutral outcome. The consequences of an action were found to play only a secondary but important role: people tended to disapprove of those actions which led to a negative outcome, even if unintentionally. This evidence was

further strengthened by the measured changes in brain activity revealing that the neural networks mostly involved in these types of moral judgments were found to be located in the temporal parietal junction (TPJ), a brain area the authors (Saxe & Kanwisher, 2003) interpret to subserve false belief judgments (the validity of this view is still under debate, a more thorough discussion on the role of the TPJ in human behavior is proposed in Section 3.1).



**Figure 4** Example of a moral scenario used by Young and colleagues (2006, figures 1a and 1b), showing how the effects of manipulating intentionality and consequences had on moral judgments

In a more recent behavioral study (Cima et al., 2010), supporters of the Universal Moral Grammar model tested yet another type of moral scenario on psychopathic patients. As discussed in the previous section, psychopaths are suggested to be unable of processing emotions (Blair et al., 2002, 1995, Damasio et al., 2000). Therefore, the aim of this study is to show that emotions do not have an influence on moral judgments by directly comparing the moral judgments reported by psychopaths against the ones reported by healthy individuals. It is important to note that this study could contradict Prinz’s argument (2007), which is based solely on indirect findings (see Section 1.3), that morality is of emotional nature given that psychopaths are not able to behave morally because they lack the capacity of processing emotions.

In this study (Cima et al., 2010) psychopaths and healthy individuals are asked to make moral judgments on the so-called trolley-dilemmas moral scenarios. These scenarios represent types of situations in which one person has the possibility to save a greater number of human



lives (usually five people) at the expense of harming a smaller number of human lives (usually one person). In order to achieve this, in one case a person has to act on an object or tool to save lives (e.g., using a switch to deviate a runaway train from its path onto a side-track where only one workman will be killed, to prevent it from killing a group of track workers, thereby acting in a utilitarian way), while in the other case the person has to act directly on another person (e.g., pushing someone from a bridge onto a track to prevent a train from killing a group of track workers) to save the lives of a greater number of people (these scenarios are more thoroughly introduced in the next section).

Consistently with the Universal Moral Grammar model, the authors of this study (Cima et al., 2010) predicted and found that the moral judgments reported by psychopathic patients did not differ with respect to those expressed by healthy controls. Therefore, it seems that a deficit in the emotional mechanisms does not affect humans' capacity of making moral judgments. It must however be noted that these results are rather weak, given that the sample used by these authors is extremely small (only ten psychopath were tested against ten healthy individuals). Furthermore, being the absence of evidence not evidence of absence, it is usually not accepted among the scientific community to derive implications from the inability of finding an effect as many other factors such as the size or characteristics of the sample investigated might have determined the failure of producing an effect.

Combining this last piece of results with the data provided for by Haidt in his first study (Haidt, 2001) showing that conscious reasoning seems not to be playing a role in determining moral judgments, Hauser claims to have sufficient evidence supporting the hypothesis that neither conscious reasoning, as evidence suggests that moral judgments result from intuitions, nor emotions, as the data at hand suggests that moral intuitions are not of emotional nature, play a role in moral judgment. Therefore, of the three models discussed above (see figure 3), Hauser claims that only the third one is consistently corroborated by all the existing evidence. Furthermore, this model is also capable of explaining why psychopaths behave in an anti-social fashion, as suggested by the mentioned studies (Blair et al., 2002, 1995, Damasio et al., 2000): although they can distinguish right from wrong, they lack the emotional motivation necessary to implement morally adequate behavior (Cima et al., 2010).

Moreover, addressing the evidence proposed so far supporting the view that moral intuitions are of emotional nature (Moll et al., 2008; Schnall et al., 2008a, 2008b; Prinz, 2006, 2007; Wheatley and Haidt, 2005; Greene et al., 2001; Haidt, 2001), Hauser and colleagues (Huebner et al., 2008) argue that

The current evidence is insufficient to support the hypothesis that emotional processes mediate our intuitive moral judgments, or that our moral concepts are emotionally constituted. (Huebner et al., 2008, p. 5).

Taken together the results defending a causal role for moral judgment showed that a) moral violations frequently induce emotional reactions leading one to feel anger, disgust, contempt or shame depending on the situation (see Rozin et al., 1999), b) manipulating the emotional state of a person influences moral judgments, and c) brain areas associated with emotions processing are often activated during moral decision-making. These results are not conclusively demonstrating that emotions are causally involved in moral judgments as, with respect to a) emotions can simply accompany moral judgments, and this does not mean that former cause or constitute the latter; with respect to b), emotions might be simply distorting our attention towards the morally relevant elements of a situation (Huebner et al., 2008, p.1); finally with respect to c), Hauser and colleagues highlight a well-known limitation of the fMRI studies, namely due to precision issues these cannot inform us regarding the causal or the temporal role of emotions in morality (Huebner et al., 2008, p.2).

Concluding this section, Hauser and colleagues (Cima et al., 2010; Hauser et al., 2008, 2006; Huebner et al., 2008; Mikhail, 2007) proposed a model according to which moral judgments result from a Moral Organ, constituted by a moral-specific neural network in our brains which is responsible of evaluating morally relevant situations and as a result of producing the corresponding moral judgment. The discussed results show that intentions, on a larger scale, and outcomes are two elements of a situation in which the Moral Organ performs its computations in order to evaluate such situations. Furthermore, he proposed some findings suggesting that emotions are not involved in the causal process leading to moral judgments showing that psychopaths and healthy individuals make similar moral judgments. Hauser's

proposal is surely interesting but not fully convincing. In fact, focusing on falsifying the evidence supporting the hypotheses that emotions play a role in moral judgments, Hauser or other supporters of the Universal Moral Grammar theory have so far not proposed any solid evidence confirming the existence of a Moral Organ, or more broadly, suggesting that in order to make moral judgments we follow a logical structure such as the one entailed by this theory. Hauser's theory has however the important merit of having stressed a crucial previously overlooked point: moral intuitions do not need to be of emotional nature.

In the section that follows I focus on Greene's dual-process theory (2004, 2002, 2001) in which two psychological mechanisms, one cognitive and one emotional, are hypothesized to be supporting moral judgments. Beforehand, however, I will discuss the origins of the moral scenarios Greene uses to collect his empirical data, since the crucial moral dilemma giving origin to these types of scenario (i.e. when is it permissible to harm someone in order to promote a better good?) have been the center of controversy in moral philosophy for several decades, if not centuries (Thomson, 1990, 1986; Foot 1978, 1967).

### **1.5. Philosophical Origins of the Trolley Dilemma**

Before describing in greater details the dual-process model I will first introduce the philosophical origins of the moral dilemma extensively used to test the predictions of this theory, namely: when is it morally permissible to harm someone?

Since such a dilemma was first addressed in Thomas Aquinas' Summa Theologica (IIa-IIae Question 64 article 7) it has been at the core of moral philosophers' attention who proposed several different theories and principles to establish which conditions must be respected in order to consider a harmful action morally permissible. For instance one of the most famous sets of principles can be found in Aquinas' doctrine of double effect (IIa-IIae Question 64 article 7):

A person may not intend evil, even when the evil will be a means to a greater good. Nonetheless, one is permitted to employ neutral or good means to promote a greater good, even though we foresee evil side effects if the good is proportionate to the evil and if there is no better way to achieve such good (Kamm, 2000, p. 211).

In recent years two branches of moral philosophy have been most influential in prescribing a set of principles prescribing moral behavior in these types of situations: utilitarianism (or consequentialism) and deontology (or non-consequentialism). By placing different emphasis on the motivation and consequences of an action, these two theories offer contrasting explanations of how to determine if an action is morally permissible or not. On one hand, supporters of the former view propose that in order to assess the permissibility of an action one should consider solely the foreseen consequences of the action: if acting in a certain way promotes the realization of the greater good, that action is morally permissible.

On the other hand, defenders of the latter position assert that consequences, despite being very important, are not the only element one should take into account to evaluate the moral permissibility of an action. In addition to the consequences of his action, one should always and more importantly establish if the action being evaluated respects one or more pre-established moral rules. The best example of those moral rules is perhaps the Kantian Categorical Imperative (Kant, 1785/1959) stating that one must act in such a way that he treats humanity, whether in his own person or in the person of another, always at the same time as an end and never simply as a means (Kamm, 2000).

At the theoretical level both positions are well supported by convincing arguments, with the result that accepting one view over the other depends on the propensity of a person to accept the meta-ethical principles behind each of these two views. Therefore, an attractive alternative to provide one or the other position with stronger support can be found in adopting a more pragmatic approach. Pursuing this alternative path the most recent works in moral philosophy concerned with this type of dilemmas, shifted their attention from the meta-ethical domain to folks' moral intuitions. In other words, the rationale behind this pragmatic approach is that in order to decide which of the two moral theories is more appropriate to guide people's actions in this type of moral decisions, philosophers should stop discussing the meta-ethical validity of the models and start concentrating on individuals' intuitions about what is morally permissible and what not. These intuitions will then be used to test which of the two theoretical frameworks, utilitarianism or deontology, is better in representing individuals' intuitions.

The introduction of studies which placed their focus on individuals' intuitions was received in multiple ways by philosophers. On one hand some (Held, 2002; Korsgaard, 1996; Ruse, 1995; Farber, 1994) dismissed these studies raising the naturalistic fallacy argument (Moore, 1903) against them. In other words these studies are rejected stating that even if these may be very useful to describe how a moral judgment happens (*is*), these have nothing to contribute to the normative debate on what the moral judgments *should* be. Further, as moral philosophy focused solely on normativity issues, pragmatic studies have nothing to contribute to the philosophical debate. In contrast, others acknowledge that such a pragmatic approach may vastly benefit moral philosophy (Lawrence & Calder, 2004; Pidgen, 1991; Mackie, 1980). For instance, as we have seen, Prinz (2007, 2006) puts individuals' intuition, and the empirical evidence gathered from studies analyzing moral behavior, at the core of his moral theory.

One of the very first contributions to moral philosophy approaching the dispute on the righteous harm dilemma from this new pragmatic approach can be found in Philippa Foot's works (1978, 1967). Here, while debating on the permissibility of abortion, the author introduces a thought experiment to defend her claims, labeled the *trolley dilemma*:

A runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save them is to hit a switch that will turn the trolley onto an alternate set of tracks where it will kill one person instead of five. Ought you to turn the trolley in order to save five people at the expense of one? (Greene et al., 2001, p. 1)

Several different forms of this dilemma have been proposed since its original appearance, for instance Judith J. Thomson (1990, 1986) proposed a slightly different one, known as the *footbridge dilemma*:

A runaway trolley is headed for five people who will be killed if it proceeds on its present course. You are standing next to a large stranger on a footbridge that spans the tracks, in between the oncoming trolley and the five people. In this scenario, the only way to save the five people is to push this stranger off the bridge, onto the tracks below. He will die if you do this, but his body will stop the trolley from reaching the others.

Ought you to save the five others by pushing this stranger to his death?  
(Greene et al., 2001, p. 1)

Despite the descriptive difference between the two alternatives of the trolley dilemma appears to be very little, it has a major effect on the response given by individuals. Indeed, the majority of individuals in fact judge that diverting the trolley is morally permissible, but the majority of them also judge it impermissible to push the man in front of the trolley (Thomson, 1990, 1986; Greene et al., 2001). Hence, from the difference in individuals' intuitions a first important statement in favor of the deontological moral theory can be inferred: consequences do not seem to be the most important element for the judgment in the second scenario, thus there must be something else taken into consideration.

The element making the difference between the two scenarios, easily derivable confronting the two of them, stands in the way in which the death of the one person is brought about: in the first case the death of the one is one of the unfortunate consequences of the action; in fact, if he was not there and one diverted the trolley, one would save five lives and no one would die; in the second case instead, one needs the man to be on the footbridge, so that one can push him in front of the trolley and *use* him to stop the trolley. If the man was not on the footbridge one could not save in any way the five. Curiously, what seems to make a big difference in determining individuals' intuitions seems to be precisely the moral principle resulting from the Kantian Imperative prohibiting treating someone as a means for one's goal (Kant, 1785/1959).

In order to support or dismiss the claim that folk's intuitions support the deontological moral theory, several additional scenarios varying some of the descriptive details have been proposed: for instance, the footbridge scenario has been modified introducing a trap-door which if opened would make the man fall in front of the trolley, so that there is no personal interaction which may distort the judgment. Other proposals include variants of the trolley, e.g. the loop track versions, or scenarios sharing a similar structure as the "transplant dilemma", where a doctor could use the organs of a healthy man in order to transplant them into five patients thereby saving their life but killing the healthy man. Up to now the debate has not been resolved yet, as it can be seen in the dispute between Frances Kamm (2000), for the deontologist view,

and Thomas Scanlon (2008), for the utilitarian side. For the aims of the present work, the most important effect brought about by the described change in abstraction level of moral philosophical research is that it catalyzed the interest of psychologists and neuroscientists for moral decision-making.

### **1.5.1. The Dual Process Theory of Moral Judgment**

In this section I will outline the last model put forward in the moral psychology literature considered in the present work, namely the Dual-Process Theory of moral judgment (Greene et al., 2004, 2002, and 2001). Proposers of this theory argue that moral judgments result from multiple psychological systems, claiming that both conscious reasoning and emotions are involved in moral judgments and suggesting that moral intuitions may rely on both affective and cognitive mechanisms. Crucially, these authors also stress the point that different moral judgments (e.g. utilitarian or deontological judgments) are underpinned by different psychological systems (Cushman et al., 2010). In other words, the observed phenomenon (see above, and Thomson, 1990, 1986) of people making different moral judgments in very similar situations may be explained by the fact that different psychological mechanisms relying on distinct neural networks are involved in the two types of moral decision making.

The scholars supporting this dual-process model concentrate their work on testing the empirical validity of the following hypotheses: a) different types of moral judgments are processed by different psychological mechanisms relying on distinct neural networks; b) these mechanisms are one emotional and the other cognitive, the former is associated with deontological moral judgments while the latter with utilitarian moral judgments (this hypothesis is largely informed by the findings gathered to support hypothesis a), and c) manipulating one of these mechanisms should predominantly affect only the moral judgments which are associated with it, and not the other ones.

In order to test the very first hypothesis Greene and colleagues (2001) used fMRI to investigate the neural correlates of moral judgments in the two moral scenarios Foot (1978) and Thomson (1986) introduced to the moral philosophy literature. As mentioned before, these two scenarios are very similar, but contain a substantial difference with respect to the means by which the goal of saving the group of people is achieved: in the former the protagonist of the

scenario operates on a lever, while on the latter the protagonist acts on a person. In his work Greene used this difference to categorize dilemmas similar to the first type as impersonal moral dilemmas, while those similar to the second type as personal. More specifically, personal dilemmas are those meeting the following

Three criteria: First, the violation must be likely to cause serious bodily harm. Second, this harm must befall a particular person or set of persons. Third, the harm must not result from the deflection of an existing threat onto a different party (Greene et al., 2004, p.1)

At the behavioral level, the results obtained in this study revealed that most of the individuals judged the protagonist action to be morally acceptable in the first scenario but not in the second, thereby replicating early observations (Thomson, 1986; Foot, 1967). More interestingly, the measured neural activity revealed that the moral decisions in the two situations correlated with different neural networks. The obtained data, in fact, revealed that brain areas in the Medial Frontal Gyrus, the Posterior Cingulate Gyrus, and the Angular Gyrus were predominantly active when respondents were making judgments about personal moral scenarios, in particular when judging whether it is impermissible to kill a person in order to save others (a deontological judgment). In contrast, the parietal lobes, bilaterally, had a more important role when respondents were making judgments about impersonal moral judgments and when judging whether it is permissible to sacrifice one life to save a group of people in the personal moral scenarios (a utilitarian judgment).

This very first evidence seems to strongly support Greene's hypothesis that different psychological mechanisms are involved in different types of moral judgments. Furthermore, a closer inspection of these data revealed a more striking finding: the brain areas reported to be mostly involved in deontological moral judgments (expressed mostly in the context of personal moral dilemmas) have been previously associated with emotion processing in several neuroscientific studies (Maddock, 1999; Reiman et al., 1997; Kosslyn et al., 1996); utilitarian moral judgments (expressed on both types of dilemma, but mostly on impersonal ones) correlated with an increase of neural activity in brain areas associated in the neuroscientific



literature with cognitive processes, involved for instance in working memory tasks (Cohen et al. 1997; Smith & Jonides, 1997).

Greene's (2001) explanation of these finding is that in the case of the impersonal moral dilemma the moral decision is mostly informed by the cognitive calculus that saving five people at expenses of one is better than the opposite outcome. However, in the case of the personal scenario, the moral decision is also influenced by a strong aversive emotional reaction towards the act of directly killing someone. The moral judgment is therefore the outcome of a competition between the cognitive information suggesting that the more desirable choice is that of saving more people against the emotional information suggesting that it is very undesirable to directly harm a human being. The mechanism which prevails over the other will determine which moral decision a person will make: if the cognitive prevails then one will judge pushing the person permissible, hence if the emotional prevails one will judge it not morally permissible.

Further, according to Greene (2004, 2002, and 2001), evidence that there are multiple mechanisms competing against each other when having to make a moral judgment in a personal type of dilemma situation is revealed by a) the reaction times required for a person to make a decision: it takes her much longer to decide in the case of personal moral dilemmas compared to impersonal ones, and b) by the activation of the anterior cingulate cortex, a brain area associated with decisions conflict resolution as for instance in the stroop task (for a review see Botvinick et al., 2001; MacLeod, 1997). This explanation was further confirmed by data gathered in a later study (Greene et al., 2004) focused precisely on identifying the neural bases of cognitive conflict and control in moral judgment.

Largely informed by this initial evidence, supporters of the Dual-Process theory develop a series of studies aimed at corroborating the hypothesis that the two mechanisms involved in moral judgments are one emotional and the other cognitive. In what follows I discuss three studies supporting this theory: the first study observes the effects of inducing a positive mood in people before asking them to express their moral judgment on both the two types of moral scenario; the second one tests these two types of moral judgments on participants with brain damages in a specific area of their frontal cortex (see below for details); and finally the third

study examines the effects on both types of moral judgments of re-directing individuals' cognitive attention to an exogenous task while making moral judgments.

Valdesolo and DeSteno (2006) used the trolley and the footbridge dilemma to test the effects of inducing a positive emotion, in this case amusement, on moral judgments. Consistently with Greene's Dual-Process theory, they predicted that inducing amusement should influence individuals' moral judgment solely when they are considering personal moral dilemmas, and not their moral decisions with respect to impersonal moral dilemmas. To obtain this evidence the authors showed to the people participating in the study, either a five minute clip from a comedy show (expected to amuse them) or a segment of a documentary (expected to not alter participants' emotional state).

They then asked them to express their moral judgment both on the footbridge scenario (a personal dilemma) and on the trolley scenario (an impersonal dilemma). Comparing the moral judgments expressed by people in a neutral emotional state with those who were amused, the authors found that being in a more positive affective state increased the odds of judging the protagonist's action morally permissible only in the personal moral scenarios. In sum, this study showed that modifying a person emotional state would only influence her moral judgments in the case of personal moral dilemmas.

In a similar vein, Koenigs and colleagues (2007) tested the same hypothesis entailing that emotions are involved only in personal moral judgments. This time however, instead of inducing emotions in people, the authors presented the personal and impersonal moral dilemmas to patients with bilateral damage to the ventromedial prefrontal cortex (vmPFC), a brain area associated with affective valuation processes (Damasio et al., 2000). Comparing the moral judgments of these patients with those of healthy individuals revealed that lesion patients reported more utilitarian judgments on personal moral dilemmas than healthy controls. In contrast, judgments expressed on impersonal moral scenarios did not differ between the two groups of individuals (Koenigs et al., 2007).

To further test the hypothesis that moral judgments in the personal moral scenarios result from a competition between emotional and cognitive processes, a third study analyzed the effect of cognitive load on moral judgments. Cognitive load refers to the effect of focusing the attention

of a person's cognitive resources on a given task in order to reduce the likelihood that these processes will be involved in a different task. For instance a typical way to focus the cognitive attention of a person on a task is to ask her to pay attention to a series of numbers presented on a computer screen and to report each time they see a particular number (e.g., number three). Therefore, in this study Greene and colleagues (2008) tested the prediction that people under increased cognitive load would report fewer utilitarian moral judgments in the case of personal moral scenarios compared to non-loaded individuals.

To obtain this evidence, a group of individuals was asked to express their moral judgments on personal and impersonal moral scenarios. While doing so, some of them had also to perform at the same time the previously described cognitive load task while the others did not. Greene and colleagues' hypotheses found only partial support from the evidence obtained. In fact, contrary to their expectations cognitive load did not influence moral judgments. However, as expected, the data did reveal that people under increased cognitive load required more time to express a utilitarian moral judgment (Greene et al., 2008). This evidence, although not as solid as the authors probably had expected, seems to support the claim that utilitarian moral judgments are relying mainly on cognitive neural processes. However, the fact that moral judgments were not affected by manipulating a person's cognitive capacities might constitute a strong argument against the view that the two mechanisms the dual-process theory links to moral judgments are competing against each other. As we will see in the final section of the present work, this data might be better explained by an alternative theory claiming that cognitive and emotional associations are not competing against each other, but are rather part of a larger valuation mechanism which integrates both the emotional and cognitive information.

In conclusion, this section introduced the dual-process theory of moral judgment. As it was discussed, proposers of this theory claim that moral judgments result from both emotional and cognitive mechanisms. More importantly, in this theory it is argued that conscious reasoning and/or intuitions play a major role in generating utilitarian moral judgments, whereas emotions do not seem to play a role in them. In contrast, deontological moral judgments seem to rely mostly on emotion-based intuitions, given that conscious reasoning and cognitive mechanisms are less involved for this type of moral judgments (Greene et al., 2004, 2001). Ironically, if this theory is correct, one could consider emotion as the key (yet secret) motivator of Immanuel

Kant's moral theory (Kant, 1785/1959) despite his declarations dictating that emotions should not have any involvement in morality (Greene, 2007).

As for all the introduced theories, the dual-process model is not perfect either. Its weakest point in my view is the claim that different moral judgments are strictly supported by either one or another psychological mechanism. In fact, despite the solid evidence suggesting that emotions are predominantly involved in deontological moral judgments, there is no strong evidence suggesting neither that a) conscious reasoning is not also involved in these types of judgments, nor that b) emotions might not be involved in utilitarian moral judgments. More specifically, with respect to the first critique (a) we are not aware of any study which tested the effect of priming one or moral principles in individuals on moral judgments. Furthermore, with respect to the latter criticism, evidence from two studies investigating the role of emotions in the personal and impersonal moral judgments using patients with deficits in emotion processing have yielded contradicting findings (compare for instance Cima et al., 2010 and Koenigs et al., 2007). More broadly, most of the literature in moral psychology focused on battling whether emotions should or not be considered among the elements giving raise to moral judgments. For this reason, no study has yet concentrated on providing a detailed account of the specific mechanisms through which emotions might affect moral judgments.

## **1.6. Rationale**

In conclusion, the extensive literature on moral decision making makes some suggestions regarding the mechanisms and brain structures that could be responsible for moral behavior. However, due to the heterogeneity of the tasks employed in the related literature, it is hard to have a clear understanding of which of the proposed models is better in capturing the actual psychological processes supporting moral decisions. Therefore, the studies proposed in the present work aim at clarifying some of the least understood, and hence most controversial, aspects of moral judgment and behavior. In the first study we focus on providing a comprehensive understanding of how and under which circumstances emotions play a role in shaping moral judgments. To achieve this goal, we propose a novel theoretical approach suggesting that emotions influence moral judgments based on their motivational dimension, and test this prediction on all the types of moral scenarios discussed so far. This approach will not

only allow to test a previously overlooked mechanism through which emotions might exert an influence on moral judgments, but also compare which of the proposed theories is better supported by the freshly gathered evidence.

Furthermore, although most of the literature agrees that some sort of cognitive mechanism is involved in moral decisions, due to the nature of the evidence it is hard to establish a causal involvement of such type of computational mechanisms in relation to moral behavior. Thus, in the second study proposed here we test the causal relation of cognitive control on moral decisions in the context of fair monetary distributions. More specifically, in this study we test the causal and functional role of the LPFC, a brain area associated with cognitive self-control (Figner et al., 2010; McClure et al., 2007), in fairness-norm compliance behavior. This study will hence allow providing for the first time solid evidence of a cognitive mechanism in the regulation of moral behavior.

Finally, our last work presented here explores a more philosophical controversy regarding the usefulness for normative moral philosophy of studies describing how moral decisions are taken. Here we address this controversy using a pragmatic approach which, we claim, reaches both into the normative and descriptive domain of morality. More specifically we focus on one hand on explaining why people do not behave in the way they ought to, while on the other hand suggesting how we may improve people to overcome the shortcomings responsible for the discrepancy between what they are morally required to do and what they actually do. Adopting such pragmatic approach will allow clearly grasping the importance of integrating normative and descriptive moral theories.

## **2. The Role of Emotions for Moral Judgments Depends on the Type of Emotion and Moral Scenario.**

(adapted from Ugazio, G., Lamm, C. and Singer, T. (2012) *Emotion*, Vol. 12 (3),579-590)

### **2.1. Abstract**

Emotions seem to play a critical role in moral judgment. However, the way in which emotions exert their influence on moral judgments is still poorly understood. This study proposes a novel theoretical approach suggesting that emotions influence moral judgments based on their motivational dimension. We tested the effects of two types of induced emotions with equal valence but with different motivational implications (anger and disgust), and four types of moral scenarios (disgust-related, impersonal, personal, and beliefs) on moral judgments. We hypothesized and found that approach motivation associated with anger would make moral judgments more permissible, while disgust, associated with withdrawal motivation, would make them less permissible. Moreover, these effects varied as a function of the type of scenario: the induced emotions only affected moral judgments concerning impersonal and personal scenarios, while we observed no effects for the other scenarios. These findings suggest that emotions can play an important role in moral judgment, but that their specific effects depend upon the type of emotion induced. Furthermore, induced emotion effects were more prevalent for moral decisions in personal and impersonal scenarios, possibly because these require the performance of an action rather than making an abstract judgment. We conclude that the effects of induced emotions on moral judgments can be predicted by taking their motivational dimension into account. This finding has important implications for moral psychology, as it points towards a previously overlooked mechanism linking emotions to moral judgments.

### **2.2. Introduction**

Over the years, the role of emotions in morality has been the source of large controversies in moral philosophy and psychology. Philosophers have debated whether we should consider our emotional reactions when defining a certain action as morally permissible or not (Hume, 1777/1960; Kant, 1785/1959), while psychologists have traditionally focused on empirical research, in particular a) whether moral judgments stem from intuitions or from conscious

reasoning, and b) which psychological processes are involved in moral intuitions (Cushman et al. 2010). Battling on these issues, scholars have neither been able to focus on providing a detailed account of the specific mechanisms through which emotions affect moral judgments nor on the contextual elements favoring the emotional involvement in the production of moral judgments. Providing a better understanding of the interaction between types of emotions with particular types of moral judgments is therefore the main objective of our paper. Furthermore, we will try to delineate some of the circumstances under which moral judgments are more likely to rely on emotional processes.

The two issues mentioned above have been at the core of the moral psychology debate in the last decade. It has been suggested that moral judgments result mainly from intuitions and that these intuitions are of emotional nature (Haidt, 2001; Prinz, 2006; Schnall et al., 2008; Wheatley & Haidt, 2005). Haidt and colleagues, for instance, proposed that moral judgments are largely influenced by our “gut feelings”, an idea previously suggested by Hume (1777/1960). To test this, Haidt and colleagues (Schnall et al., 2008; Wheatley & Haidt, 2005) developed a series of moral vignettes describing violations of moral norms that are strongly connected to feelings of disgust. For instance, one of the most representative scenarios describes two siblings who decide to have sexual intercourse. Haidt hypothesized that such a scenario would induce a feeling of disgust in the participants and that this feeling would influence the outcome of their judgment concerning whether the protagonists’ intention is morally permissible. Different techniques were used to induce disgust in several studies, including hypnosis (Wheatley & Haidt, 2005) and disgusting smells (Schnall et al., 2008). These studies showed that the induction of disgust led to more severe moral judgments, supporting the hypothesis that moral judgments are linked to primary emotions, in this case, disgust. Furthermore, Schnall et al. (2008) showed that the effect on moral judgments was specific for disgust and not for other emotions, such as sadness, indicating that disgust is specifically linked to moral judgments related to scenarios that might trigger disgust. For this reason, these types of moral scenarios are labeled *disgust-related* in the present paper.

An alternative theory agrees that moral judgments result from intuitions, but denies any role for emotions in the formation of moral judgments, claiming that moral intuitions result from a moral specific psychological mechanism labeled “universal moral grammar” which focuses

mostly on the protagonist's intentions in order to determine whether an action is permissible or not (Hauser, 2006; Huebner et al., 2008; Mikhail 2007). More importantly, the "moral grammar" (for a critique of this theory, see Dupoux & Jacob 2007) is independent of both emotional and cognitive mechanisms. According to this view, emotional and cognitive processes are typically activated *after* a moral judgment has been made. Emotions and reasoning thus have no causal role in determining moral judgments (Hauser, 2006; Huebner et al., 2008). Using personal and impersonal moral scenarios similar to those introduced by Greene et al. (2001) in the moral psychology literature, Cima et al. (2010) provided the strongest evidence against the claim that emotions are related to moral judgments by showing that psychopaths' moral judgments do not differ from those of healthy participants. This has been interpreted as contradictory to the claim that emotions play a role in determining moral judgments, as psychopaths are known to have impaired affective processing (Blair et al., 2005).

To further test the hypothesis that it is the protagonist's intention which determines moral judgments, another type of scenario was introduced (Young et al., 2006, 2010). These scenarios, labeled *beliefs* scenarios here, describe a person knowingly (versus unknowingly) causing a negative outcome (e.g., someone's death), that is, believing that his actions will (versus will not) lead to the negative outcome. The neural networks mostly involved in these types of moral judgments were found to be located in the temporal parietal junction (TPJ), a brain area the authors interpret to subserve false belief judgments (Saxe et al., 2003). In contrast to such a narrow view of a specific role of TPJ in Theory of Mind processing, a more frequent view is that TPJ is more broadly involved in general processes related to mentalizing, detecting multi-sensory integration incongruency (e.g. visual and proprioceptive signals from the body), and orienting attention away from self-related processing (Apperly et al. 2007; Decety & Lamm 2007; Mitchell 2005).

From a third point of view, some scholars try to reconcile the contradictory positions mentioned above. Greene and colleagues proposed a dual-process theory for moral judgments (Greene et al., 2004). They argue in this theory that moral judgments result from multiple psychological systems, claiming that both conscious reasoning and emotions are involved in moral judgments and suggesting that moral intuitions may rely on both affective and cognitive



mechanisms. Furthermore, these authors also stressed the point that different moral judgments are underpinned by different psychological systems (Cushman et al. 2010).

Greene et al. (2001, 2004) used functional magnetic resonance imaging (fMRI) to test this theory and to investigate the neural correlates of moral judgments in two moral scenarios Thomson (1986) introduced to the moral philosophy literature. These scenarios represent types of situations in which one person has the possibility to save a greater number of human lives (usually five people) at the expense of harming a smaller number of human lives (usually one person). In order to achieve this, in one case a person has to act on an object or tool to save lives (e.g., using a switch to deviate a runaway train from its path onto a side-track where only one workman will be killed, to prevent it from killing a group of track workers, thereby acting in a utilitarian way), while in the other case the person has to act directly on another person to save the lives of a greater number of people (e.g., pushing someone from a bridge onto a track to prevent a train from killing a group of track workers). Thus, the scenarios differ with respect to the means by which the goal of saving the group of people is achieved. Typically, this difference leads respondents preferring utilitarian judgments in the first scenario, later labeled *impersonal*, while being less utilitarian in their judgments in the second scenario, labeled *personal* (Greene et al., 2001). The neuroimaging data obtained by Greene and colleagues (2001, 2004) suggested that emotional processes were predominantly active when respondents were making judgments about *personal* moral scenarios, in particular when judging whether it is impermissible to kill a person in order to save others. In contrast, cognitive mechanisms played a more important role when respondents were making judgments about *impersonal* moral judgments and when judging whether it is permissible to sacrifice one life to save a group of people in the *personal* moral scenarios (Greene et al., 2001, 2004, 2008). These results thus show that cognitive mechanisms inform moral judgments based predominantly on an action's consequences (utilitarian judgments), while emotional mechanisms primarily inform moral judgments focusing on the means used to obtain a given outcome (deontological judgments).

In line with these conclusions, several other studies suggest that emotions are indeed strongly involved in personal and less involved in impersonal moral judgments. For example, a behavioral study by Valdesolo and DeSteno (2006), using one personal and one impersonal moral scenario, showed that being in a more positive affective state increased the odds of judging

the protagonist's action morally permissible only in the personal moral scenarios. Moreover, lesion patients in a study involving patients with bilateral damage to the ventromedial prefrontal cortex (vmPFC), a brain area associated with affective valuation processes, judged the actions of protagonists in the personal moral scenarios, but not in the impersonal scenarios, to be morally permissible significantly more often than normal controls (Greene, 2007; Koenigs et al., 2007; Young & Koenigs, 2007). Note however, that the interpretation of the Koenigs et al. findings is not univocal. In fact, while Moll and colleagues (2007) acknowledge that emotions are predominantly involved in personal scenarios, they propose an alternative explanation of these findings, suggesting that the moral judgments of patients with vmPFC lesion differ from those of healthy subjects because the former are unable (or less able) to experience prosocial moral sentiments. They are for this reason more prone to judge moral dilemmas in a more utilitarian fashion and not because they are unable to process emotionally salient information, as Greene (2007) suggests.

Moll and colleagues' interpretation is derived from a fourth theory on moral judgments which suggests that both emotions and conscious reasoning are involved in moral judgments, similarly to Greene's theory. Contrary to the latter, however, which holds that some moral judgments can only entail emotional or cognitive processes, Moll's position suggests that all moral scenarios will involve cognitive and emotional associations competing against each other at the moment of producing the corresponding moral judgment (Moll et al., 2005, 2008).

As the brief literature review makes clear, most of the research investigating the role of emotions for moral judgments focused primarily on the question of whether induced emotion has an effect on moral judgments. Empirical evidence has provided evidence for a role of emotions in moral judgments, speaking against those theories stating no causal role in moral judgments (Cima et al., 2010; Hauser et al., 2006; Huebner et al., 2008). Recently, some attempts have also tried to reconcile the contra posed views (Cushman et al., 2010). The main reasons for these contradictory findings and disagreement can be found in the fact that no study has systematically compared the interaction between different types of emotions on different types of moral dilemma. The present study aims at taking a first step in filling this gap. To do so, we focused on the motivational dimension of emotions and their differential role on moral judgments typically used in the literature.

Recent research on social cognition has indicated that the effects of emotions on behavior can be better understood if emotions are classified according to their motivational dimension rather than their valence (Forgas, 2003; Harlé & Sanfey, 2010). More specifically, researchers investigating the relationship between motivation and emotion suggest classifying emotions according to approach and withdrawal motivational tendency rather than valence alone (Berkowitz, 2003; Harmon-Jones, 2004; Lang et al., 1997; Spielberg et al., 2008). Approach emotions are those which are more likely to result in behavior involving approaching a certain person, situation or event, while withdrawal emotions rather result in the opposite, namely, withdrawal of persons, situations, or events. Although other dimensions of emotions may be taken into consideration in order to predict their effects on moral decision making (Power and Dalgleish, 2008), we chose the motivational tendency as: a) in order to express their moral judgment on a given behavior, subjects are required to imagine the described action and have to decide whether they would endorse it or not, b) previous research on social decision making has mainly focused on the dimensions of motivational direction and valence. Harle and Sanfey (2010), for instance, predicted differential effects of emotions based on the motivational direction dimension, while others focused on the valence dimension (Schnall et al., 2008; Valdesolo and DeSteno, 2006; and Wheatley and Haidt, 2005). These studies provide a sufficient foundation for making differential predictions regarding how emotions would affect moral judgments if one or the other dimension prevails. In fact, testing two emotions with shared negative valence but opposite motivational tendencies (anger and disgust) allows us to formulate mutually exclusive predictions on the effect emotions one should expect on moral judgments (as described in detail below).

To unveil whether emotions' motivational dimension is crucial in determining differential effects on moral judgments, the present study was designed to investigate how emotions differing in motivational tendency influence moral judgments and how these effects depend on the type of moral scenario being judged. More specifically, we tested the effects of two primary negative emotions with equal valence but opposite motivational directions: anger (approach motivation) and disgust (withdrawal motivation). Furthermore, to assess whether the effect of induced emotions also depends on the type of moral scenario, this study directly compared moral judgments expressed on four types of moral scenarios. This entailed the inclusion of scenarios

where we did not expect emotions to exert an effect on moral judgments. To this end, we used the four types of scenarios most commonly used in the literature and introduced in detail above: disgust-related moral vignettes (Haidt, 2001), impersonal and personal moral scenarios (Greene et al., 2001), and beliefs moral scenarios (Young et al., 2006). As participants were asked to evaluate whether a protagonist's action was permissible in the scenarios, we predicted a differential effect of emotions on moral judgment depending on their motivational direction (approach vs. withdrawal). More specifically, given that approach emotions motivate one to engage in a physical interaction with someone else, approach emotions were expected to increase the number of permissibility judgments. In contrast, disgust should have the opposite effect, as it primes withdrawal action tendencies, resulting in a tendency to refrain from acting on a third party and leading to a decrease in judgments of moral permissibility. Thus, based on our main hypothesis that motivational direction rather than valence determines the effect of emotion on moral judgment, we predicted that anger and disgust – although both negative emotions – would have opposite effects on moral judgments. If, on the other hand, valence alone were the main determinant of the effects of induced emotions on moral judgments, we would expect to see similar effects for anger and disgust.

Furthermore, we predicted that the observed effects on moral judgments should not only vary as a function of the type of emotion induced but also as a function of the type of moral scenario used. Based on the literature reviewed above, we predicted that emotions would have a stronger influence on moral judgments in the disgust-related and personal scenarios as compared to the impersonal or beliefs scenarios. Furthermore, disgust induction should have an especially strong effect on the disgust-related paradigms if the assumption that these judgments recruit general mechanisms evolved to process primary disgust in the human brain and body (Haidt, 2001) is correct. In contrast, emotions should not have an effect on moral judgments resulting from beliefs scenarios, as these mostly rely on inferences about abstract beliefs and about the protagonist's intentions in the scenario. Accordingly, brain areas such as the temporo-parietal junction (Young et al., 2006, 2010), which have been shown to subserve these types of inferences, are held to be part of domain-general processes (such as detecting incongruency or re-orienting attention; Decety & Lamm 2007). In contrast, brain structures directly associated with affective processing such as the insular cortex or the amygdala (Lamm & Singer, 2010;

Singer, 2006), as well as the superior temporal sulcus which has been associated with emotional processing in moral judgments (Moll et al., 2005) do not seem to play a role in making belief inferences. Finally, previous findings suggest that emotions should affect personal more than impersonal scenarios (Greene et al., 2001, 2004; Koenigs et al., 2007; Valdesolo & DeSteno, 2006; Young & Koenigs, 2007).

### **2.3.Methods**

We performed two experiments to test our predictions. Experiments 1a and 1b examined the effects of induced disgust on participants' moral judgments assessed using the four types of moral scenarios mentioned above. In Experiment 1a, disgust was induced using a disgusting odor applied with a commercially available odor dispenser as in Schnall et al. (2008) ("fart spray" consisting of ammonium sulfide in water solution, which, when sprayed, results in a disgusting odor). As the results obtained in Experiment 1a did not replicate previous findings suggesting a priming effect of disgust induction on moral judgments (Schnall et al., 2008a; Wheatley & Haidt, 2005), we performed another experiment using a different method for inducing disgust in Experiment 1b, where we used video clips with disgusting content, previously effectively used to induce disgust in an fMRI experiment performed by Harrison et al. (2007). In Experiment 2, we tested the effects of anger on moral judgments in the same types of moral scenarios used in Experiments 1a and 1b.

#### **Experiments 1a and 1b**

##### *Participants*

Fifty-five undergraduate students took part in Experiment 1a (disgust induced via odor). After providing informed consent, participants were randomly assigned to one of two conditions: an experimental condition in which the emotion of disgust was induced (group disgust/odor,  $n = 30$ , 22 females) and a control condition where no emotion was induced (group disgust/odor-control,  $n = 29$ , 23 females). One hundred and nine undergraduate students participated in Experiment 1b (disgust induced via video clip). As in Experiment 1a, participants were randomly assigned to one of two conditions: an experimental condition where disgust was induced (group

disgust/video,  $n = 56$ , 42 females) and a control condition (group disgust/video-control,  $n = 53$ , 38 females). Participants received 20 Swiss francs for participating in these experiments.

### *Procedure*

#### *Emotion Induction*

In Experiment 1a, as in Schnall et al. (2008), disgust was induced by means of “fart sprays”. Before each experimental session, two consecutive sprays were applied to a trash bag placed below the desk at which participants sat during the experiment. Participants assigned to the control condition where no emotion was induced performed the experiment in an identical room but with a neutral ambient odor. In Experiment 1b, disgust was induced in the experimental group using a 2-minute video clip showing an actor interacting with human vomit (Harrison et al., 2007). The control group viewed a neutral 2-minute video clip of a person describing a painting.

#### *Moral Judgments*

Participants read 40 moral scenarios (translated into German), 10 per scenario type: disgust-related, impersonal, personal, and beliefs. For each scenario, they were to answer the question: “Is it morally permissible for the protagonist to do x,” to which they answered *yes* or *no* by pressing one of two buttons on the computer keyboard. These scenarios were taken from previous studies: disgust-related moral scenarios from Schnall et al. 2008, personal and impersonal ones from Greene et al. 2008 and 2009, and belief moral scenarios from Young et al. 2006. Of the belief moral scenarios, four described neutral intentions resulting in three neutral outcomes and one negative one, and the other six describing bad intentions, three resulting in neutral outcomes and three in negative ones. The order of presentation of the types of scenarios was randomized across subjects to exclude any presentation order effects on moral judgments.

#### *Manipulation Checks*

At the end of the study, participants reported on an 11-point scale (from *not at all* to *very strongly*) if and how strongly, after the induction, they felt any of the following emotions: anger, disgust, happiness, or sadness. Additionally, to assess possible differences in the decay of

emotions over time, participants also reported how strongly they felt these emotions immediately before and after having responded to each moral scenario. Furthermore, participants were also asked to report whether and how strongly the emotion they felt might have influenced their moral judgments in order to assess possible effects of beliefs about the effect of emotions on behavior.

In order to increase the sensitivity of our analyses, we used the results of these rating scales to exclude participants for whom the emotion induction did not work satisfactorily, that is, who did not show the expected affective responses intended by the emotion induction procedures. To be sure that the respective emotion had been successfully induced, participants of a given experimental group had to report an emotion intensity of five or higher (i.e., above the midpoint on the 11-point scale ranging from *not at all* to *very strongly*) for the target emotion (i.e., disgust for the disgust groups, anger for the anger group, see Experiment 2). Participants of the control groups were included in the final analyses only if they reported emotion intensities of below five for all four emotions. In order to fully disclose the data obtained in our study, however, results will be reported for both the selected and the full sample.

## **Experiment 2**

### *Participants*

One hundred and twenty-two undergraduate students took part in Experiment 2 (anger induced via negative feedback). After providing informed consent, participants were randomly assigned to one of two conditions: an experimental condition in which the emotion of anger was induced (group anger,  $n = 59$ , 23 females) and a control condition in which no emotion was induced (group anger-control,  $n = 63$ , 28 females). As in Experiments 1a and 1b, participants received 20 Swiss francs for participating in the experiment.

### *Procedure*

### *Emotion Induction*

Giving negative feedback on essays written by participants is a well-established technique for inducing anger (Harmon-Jones & Sigelman, 2001). The technique proposed

originally in the mentioned study was slightly modified. First, the participants actually met the other participants who corrected their essays at the beginning of the experiment. Second, every participant had to both write and correct an essay to increase the credibility of the procedure. Furthermore, as pretests had shown that feedback interpreted as neutral was hard to obtain using the original scale, the nine-point feedback scale used in the original version was modified to a scale with only three evaluation categories (negative, neutral, positive).

Participants were given ten minutes to write a short essay discussing one of five controversial topics. One topic, for example, was “should alcoholic drinks be sold to people under 16 years of age.” Participants were allowed to decide which topic they wanted to discuss in their essay. After writing their essay, each participant received an essay written by another participant and evaluated it by giving either a negative, neutral, or positive evaluation based on four criteria: rationality, logic, interest, and intelligence. The experimenter then replaced these evaluation forms with pre-prepared evaluation forms used to induce anger, or no emotion, and gave them to the participants. Depending on their group assignment, participants received evaluations that were either negative or neutral on all four criteria. Please note that this procedure may have induced some complex social emotions such as shame or guilt, which were not controlled for given that our focus here were the effects of induced primary emotions on moral judgments.

#### *Moral Judgments and Manipulation Checks*

The procedures for moral judgments and manipulation checks were identical to those used in Experiments 1a and 1b.

#### *Statistical Analysis*

The moral judgment data were analyzed using a mixed-design analysis of variance (ANOVA), with the between-subjects factors Group (two factor levels: *experimental group*, in which emotions had been induced, and *control group*, in which no emotions had been induced) and Emotion (two levels: Experiments 1a and b with the emotion of disgust; Experiment 2 with the emotion of anger), and the within-subjects factor Type of scenario (4 factor levels: disgust-related, impersonal, personal, and beliefs). The dependent variable was the sum of permissible



responses given for each type of scenario. Data from the experimental and control groups of Experiments 1a and 1b were analyzed together after ensuring that the two groups did not differ significantly (see results).

The self-report data of the manipulation checks were analyzed as follows: In order to assess whether we had successfully induced emotions in the experimental groups, an ANOVA with the between-subjects factors Group and Emotion and the within-subjects factor Affect Rating (4 factor levels: anger, disgust, happiness, sadness) was performed. To establish whether anger and disgust intensity decayed over time and whether they differed with respect to decay, an ANOVA with the between-subjects factor Experimental Group (2 factor levels: disgust group and anger group) and the within-subjects factors Affect Rating (2 factor levels: induced disgust and induced anger) and Decay (2 factor levels: induced emotion at beginning, induced emotion at end) was performed. Finally, to test whether participants' beliefs about the effects of an induced emotion on their moral judgments varied across emotions, we performed an ANOVA with the between-subjects factor Experimental Group and the within-subjects factor Influence (2 levels: influence of disgust on moral judgments and influence of anger on moral judgments). Violations of the sphericity assumption in ANOVA omnibus tests were corrected using the method proposed by Greenhouse and Geisser. A-priori and post-hoc hypotheses of specific differences between factor levels were assessed using linear contrasts computed with specific error variance terms (Boik, 1981).

More specifically, a first set of planned comparisons assessed whether the emotion inductions were successful, as indicated by the affect ratings of the manipulation checks. To this end, linear contrasts assessed whether the ratings of the target emotion (disgust in Experiments 1a and 1b, anger in Experiment 2) were higher than the mean ratings of the non-target emotions (happiness, sadness, and anger for Experiments 1a and 1b; happiness, sadness, and disgust for Experiment 2). These contrasts were calculated separately for the experimental and control group of Experiments 1a and 1b (pooled) and Experiment 2. To assess whether the intensity of the target emotions (disgust and anger) was equal in the corresponding experimental groups, another linear contrast compared the self-reported intensities (Anger vs. Disgust: Target Emotion Intensity). Similarly, a planned comparison contrasted the self-reported beliefs of the influence of the target emotions on moral judgments (Anger vs. Disgust: Influence) in order to assess the

effect of participants' beliefs about the influence of the induced emotions on their moral judgments.

For moral judgments, a linear contrast was computed which tested whether the differences between the experimental groups and their respective control groups differed significantly, irrespective of the type of scenario (formally, corresponding to a contrast (Anger vs. Anger/Control) vs. (Disgust vs. Disgust/Control): mean of all scenarios) to assess whether the Group\*Emotion interaction was driven by anger increasing judgments of moral permissibility and disgust decreasing them. Another linear contrast tested whether the induced emotions affected the disgust-related and the personal scenarios more than the impersonal and beliefs scenarios. This contrast compared the difference in judgments of moral permissibility between the two experimental groups (compared to their respective controls) for the disgust-related and personal scenarios with that difference in the impersonal and beliefs scenarios (i.e., (Anger vs. Anger/Control) vs. (Disgust vs. Disgust/Control): (Disgust-Related and Personal vs. Impersonal and Beliefs). Furthermore, given that our data revealed a different pattern of results, we also tested whether emotions affected the impersonal and the personal scenarios more than the disgust-related and beliefs scenarios. This contrast compared the difference in judgments of moral permissibility between the two experimental groups (compared to their respective controls) for the impersonal and personal scenarios with the difference in the disgust-related and beliefs scenarios (i.e., (Anger vs. Anger/Control) vs. (Disgust vs. Disgust/Control): (Impersonal and Personal vs. Disgust-Related and Beliefs).

In addition, we specifically test whether emotions have an effect on the beliefs scenarios, using another planned comparison that contrasted responses given in the beliefs scenarios, separately for the two emotion inductions ((Anger vs. Anger/Control): Beliefs and (Disgust vs. Disgust/Control): Beliefs). Finally, the last *a priori* contrast tested if induced disgust had an effect on the disgust-related moral scenarios (Disgust vs. Disgust/Control: Disgust-Related). ANOVAs were performed using SPSS (SPSS Statistics version 17.0) and linear contrasts using Statistica (Stata Soft Statistica version 7).

## 2.4.Results

Applying the exclusion criteria for emotion induction ratings explained above, a sample of 232 participants (143 females; 58 participants excluded) was selected for the analyses, composed of  $n = 56$  participants (44 females; 30 participants excluded) in the two disgust groups,  $n = 80$  participants in the disgust/control groups (58 females; 2 participants excluded),  $n = 42$  participants in the anger group (18 females; 17 participants excluded), and  $n = 54$  participants (23 females; 9 participants excluded) in the anger/control group. As an ANOVA with the between-subjects factor Experiment (2 levels: Experiment 1a and Experiment 1b) and the within-subjects factor Type yielded no between-subjects differences ( $F < 1$  for the main effect Experiment,  $F(3, 402) = 39.918$ ,  $p < 0.001$  for the main effect Type, and  $F < 1$  for the interaction Experiment\*Type; full sample: main effect Experiment  $F < 1$ ,  $F(3, 492) = 47.486$ ,  $p < 0.001$  for the main effect Type, and  $F < 1$  for the interaction Experiment\*Type), data from the experimental and control groups in Experiments 1a and 1b were treated together.

### *Emotion Induction*

	Selected Sample			Full Sample		
ANOVA results	F-value (df)	P Value	$\eta_p^2$	F-value (df)	P Value	$\eta_p^2$
Group	$F(1, 228) = 466.002$	$< 0.001$	0.671	$F(1, 282) = 177.562$	$< 0.001$	0.386
Emotion	$F(1, 228) = 38.915$	$< 0.001$	0.146	$F(1, 282) = 34.921$	$< 0.001$	0.110
Affect Rating	$F(3, 684) = 92.928$	$< 0.001$	0.290	$F(3, 846) = 92.928$	$< 0.001$	0.290
Group*Emotion	$F(1, 228) = 10.03$	0.002	0.042	$F(1, 282) = 2.199$	0.139	0.008
Group*Affect Rating	$F(3, 684) = 103.521$	$< 0.001$	0.312	$F(3, 846) = 43.402$	$< 0.001$	0.133
Emotion*Affect Rating	$F(3, 684) = 150.857$	$< 0.001$	0.398	$F(3, 846) = 93.818$	$< 0.001$	0.250
Group*Emotion*Affect Rating	$F(3, 684) = 166.228$	$< 0.001$	0.422	$F(3, 846) = 79.525$	$< 0.001$	0.220

**Table 1:** Summary of the ANOVA results for the self-reported emotion ratings.

As expected, all main effects (Group, Emotion and Affect Rating) as well as the interactions (Group\*Emotion and Group\*Emotion\*Affect Rating) were significant. Furthermore, the interaction Affect Rating\*Emotion and Affect Rating\*Group were also significant. Table one above provides the full statistical analyses details. An overview of the self-rated emotion intensities for the selected samples can be found in Figure 5 below.

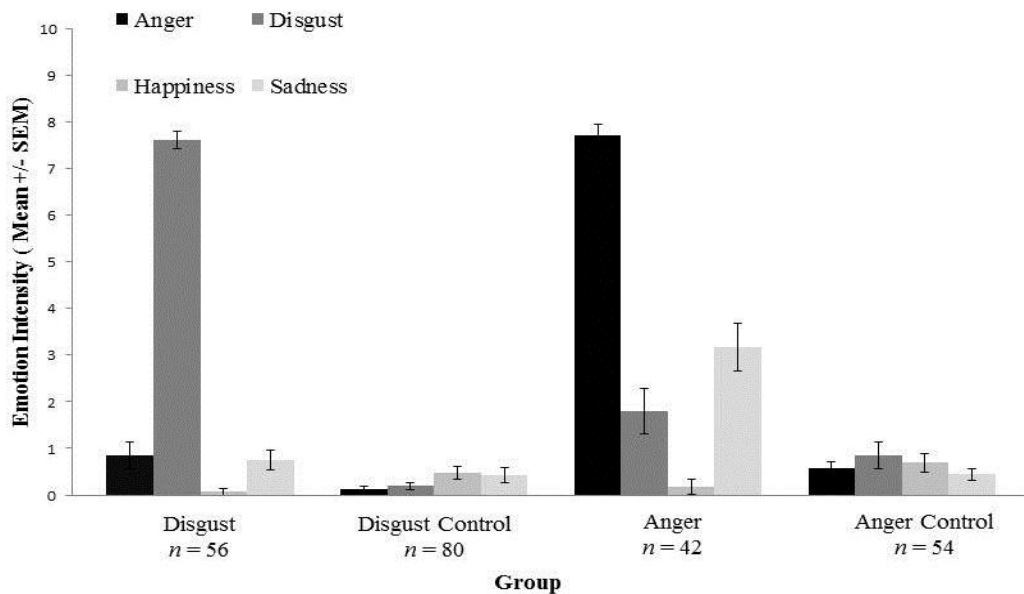
Both methods used to induce disgust in Experiments 1a and 1b resulted in high average ratings of disgust in the full non-selected sample (Experiment 1a:  $M = 4.76$ ,  $SEM = .443$ ; Experiment 1b:  $M = 6.64$ ,  $SEM = .354$ ) as well as in the selected sample; the latter were obviously higher (Experiment 1a:  $M = 6.73$ ,  $SEM = .228$ ; Experiment 1b:  $M = 7.93$ ,  $SEM = .235$ ). Therefore, the analyses that follow on the data obtained in Experiments 1a and 1b are collapsed (Experiments 1a and 1b:  $M = 7.61$ ,  $SEM = .195$ ; full sample:  $M = 5.99$ ,  $SEM = .293$ ).

Linear contrasts	Selected Sample		Full Sample	
	F-value (df)	P Value	F-value (df)	P Value
D vs. DC: d.i.	$F(1,228) = 531.837$	$< 0.001$	$F(1,282) = 263.497$	$< 0.001$
D vs. DC: a.i., h.i., s.i.	$F(1,228) = 1.753$	0.187	$F < 1$	-
D: d.i. vs. a.i., h.i., s.i.	$F(1,228) = 631.186$	$< 0.001$	$F(1,282) = 430.289$	$< 0.001$
DC: d.i. vs. a.i., h.i., s.i.	$F < 1$	-	$F < 1$	-
DC: a.i. vs. d.i., h.i., s.i.	$F(1,228) = 2.04$	0.155	$F(1,282) = 1.018$	0.314
DC: h.i. vs. d.i., a.i., s.i.	$F(1,228) = 1.731$	0.19	$F < 1$	-
DC: s.i. vs. d.i., a.i., h.i.	$F < 1$	-	$F < 1$	-
A vs. AC: a.i.	$F(1,228) = 638.529$	$< 0.001$	$F(1,282) = 153.656$	$< 0.001$
A vs. AC: d.i., h.i., s.i.	$F(1,228) = 1.753$	0.187	$F(1,282) = 1.006$	0.317
A vs. AC: s.i.	$F(1,228) = 44.485$	$< 0.001$	$F(1,282) = 42.826$	$< 0.001$
A: a.i. vs. d.i., h.i., s.i.	$F(1,228) = 644.232$	$< 0.001$	$F(1,282) = 256.7$	$< 0.001$
A: a.i. vs. s.i.	$F(1,228) = 141.415$	$< 0.001$	$F(1,282) = 80.142$	$< 0.001$
AC: a.i. vs. d.i., h.i., s.i.	$F(1,228) = 2.04$	0.155	$F(1,282) = 3.19$	0.075
AC: a.i. vs. a.i., h.i., s.i.	$F < 1$	-	$F < 1$	-
AC: h.i. vs. a.i., a.i., s.i.	$F(1,228) = 1.731$	0.19	$F(1,282) = 1.29$	0.256
AC: s.i. vs. a.i., a.i., h.i.	$F < 1$	-	$F < 1$	-
A: a.i. vs. D: d.i.	$F < 1$	-	$F < 1$	-

**Table 2:** Summary of the statistics for analyses of the self-reported emotion ratings assessed by means of linear contrasts. Capital letters abbreviations A, AC, D, and DC refer to the Anger, Anger/Control, Disgust, and Disgust/Control groups respectively. Non-capital letters abbreviations a.i., d.i., h.i., and s.i. refer to induced emotion intensity, respectively: anger intensity, disgust intensity, happiness intensity, and sadness intensity.

Linear contrasts (statistics reported in detail in Table 2) revealed the following findings: The linear contrast comparing self-reported values of disgust intensity showed that disgust was

felt significantly more strongly in the disgust groups than in their respective control groups. Furthermore, the linear contrast comparing the average ratings of the non-target emotion intensities (anger, happiness, and sadness) did not reveal any significant differences. Additionally, disgust was felt significantly more strongly than the average of the other non-target emotions in the experimental group, as revealed by a linear contrast comparing disgust intensity against the non-target emotions' intensities. The disgust/control group was successfully kept in a neutral emotional state as shown by a linear contrast comparing the intensity of each emotion with the mean intensities of the other emotions.



**Figure 5** Self-reported emotion intensity (M +/- SEM) for the selected participants in the experimental and control groups. Ratings were provided on a scale from 0 (*not at all*) to 10 (*extremely*) in response to the question “How strongly did you feel this emotion?”

In a similar vein, Experiment 2 revealed that anger was felt more strongly in the experimental group than in the control group, and the target emotion in the experimental group was felt significantly more strongly than the average of the non-target emotions. Note, however, that an examination of Figure 5 suggested that participants in the experimental group felt moderate levels of sadness. A post-hoc linear contrast computed to assess this observation revealed that sadness was significantly stronger in the anger group than in the control group.

More importantly, a further linear contrast revealed that self-reported anger intensity in the experimental group was significantly higher than sadness. Comparing the self-rated intensities of emotions in the anger/control group revealed that no emotion was felt significantly more strongly than the average of the others. Finally, comparing emotion intensities in the experimental groups across the two experiments revealed that there was no difference in the intensities of the target emotions (disgust and anger). Analyses of the decay of the target emotions revealed that emotions were felt more strongly at the beginning than at the end of Experiments 1a, 1b, and 2 (Decay:  $F(1, 96) = 91.797, p < 0.001, \eta_p^2 = 0.489$ ; full sample:  $F(1, 141) = 81.763, p < 0.001, \eta_p^2 = 0.367$ ). However, the main effects Experimental Group ( $F(1, 96) = 3.212, p = 0.076, \eta_p^2 = 0.32$ ; full sample:  $F(1, 141) = 2.153, p = 0.144, \eta_p^2 = 0.15$ ), and Affect Rating ( $F(1, 96) = 2.036, p < 0.156, \eta_p^2 = 0.014$ ; full sample:  $F(1, 141) = 2.036, p = 0.156, \eta_p^2 = 0.014$ ) and the interaction Affect Rating\*Decay ( $F < 1$ ; Full sample:  $F < 1$ ) did not reveal any significant differences. The interaction Experimental Group\*Decay was not significant either ( $F < 1$ ; full sample:  $F < 1$ ), while the interaction Experimental Group\*Affect Rating\*Decay was significant ( $F(1, 96) = 93.384, p < 0.001, \eta_p^2 = 0.493$ ; full sample  $F(1, 141) = 100.453, p < 0.001, \eta_p^2 = 0.416$ ). The linear contrast (anger group vs. disgust group) vs. (induced emotion beginning vs. induced emotion end) showed that the two target emotions did not differ with respect to decay ( $F < 1$ ; full sample  $F < 1$ ). Finally, no general effect of the belief about the influence of induced emotions on moral judgments was revealed (main effect Influence:  $F(1, 96) = 1.632, p = 0.205, \eta_p^2 = 0.017$ ; full sample:  $F < 1$ , and Experimental Group:  $F < 1$ ; full sample:  $F(1, 141) = 1.176, p = 0.280, \eta_p^2 = 0.008$ ). However, the interaction Experimental Group\*Influence ( $F(1, 96) = 37.566, p < 0.001, \eta_p^2 = 0.281$  (Full Sample  $F(1, 141) = 40.068, p < 0.001, \eta_p^2 = 0.221$ )) was significant, reflecting that participants reported a slightly stronger influence of the induced target emotion on their moral judgments in the anger group than in the disgust group. The influence ratings, however, were comparably low for both groups (disgust influence  $M = 2.33$  SEM = 0.401 and anger influence  $M = 1.77$  SEM = 0.272; full sample: disgust influence  $M = 1.57$  SEM = 0.217 and anger influence  $M = 1.90$  SEM = 0.312), and their difference might also reflect presumed differences in influence explained *ex post*, rather than actual influence during judgments.

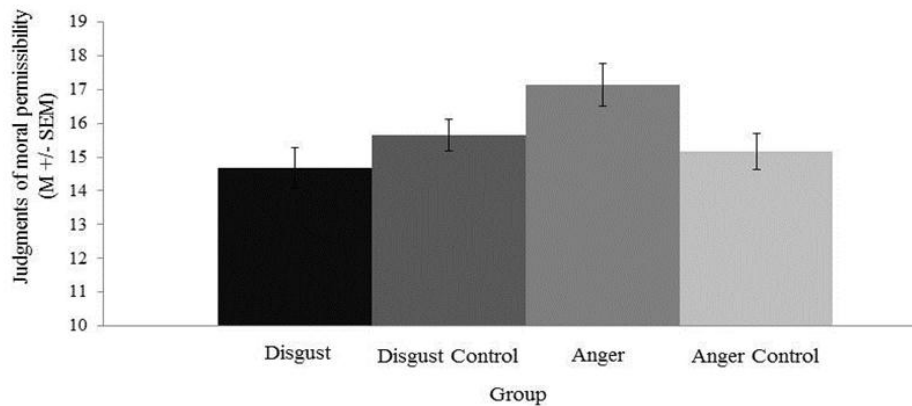
## Moral Judgments

As in the Emotion Induction results section, the full details of the statistical analyses are presented in Tables 3 and 4. Table 3 describes the ANOVA results (Group, Type, and Emotion), while Table 4 shows results of the linear contrasts. The main effects of Group and Emotion, and the interactions Group\*Type and Type\*Emotion, were not significant.

ANOVA results	Selected Sample			Full Sample		
	F-value (df)	P Value	$\eta_p^2$	F-value (df)	P Value	$\eta_p^2$
Group	$F < 1$	-	-	$F(1,282) = 2.38$	0.124	0.080
Emotion	$F(1,228) = 2.95$	0.085	0.130	$F(1,282) = 3.983$	0.047	0.140
Type	$F(3,684) = 70.428$	$< 0.001$	0.236	$F(3,846) = 98.269$	$< 0.001$	0.258
Group*Emotion	$F(1,228) = 6.989$	0.009	0.030	$F(1,282) = 2.199$	0.139	0.008
Group*Type	$F < 1$	-	-	$F < 1$	-	-
Emotion*Type	$F(3,684) = 1.756$	0.164	0.080	$F(3, 846) = 1.108$	0.340	0.040
Group*Emotion*Type	$F(3, 684) = 2.45$	0.073	0.011	$F(3,846) = 2.218$	0.095	0.008

**Table 3:** Summary of the ANOVA results for the analyses of moral judgments.

The linear contrast assessing whether anger resulted in more judgments of moral permissibility than disgust was significant (see Table 4, first row). Further planned comparisons revealed that, contrary to our expectations, the disgust-related and personal scenarios did not differ significantly from the impersonal and beliefs scenarios (see Table 4, second row).



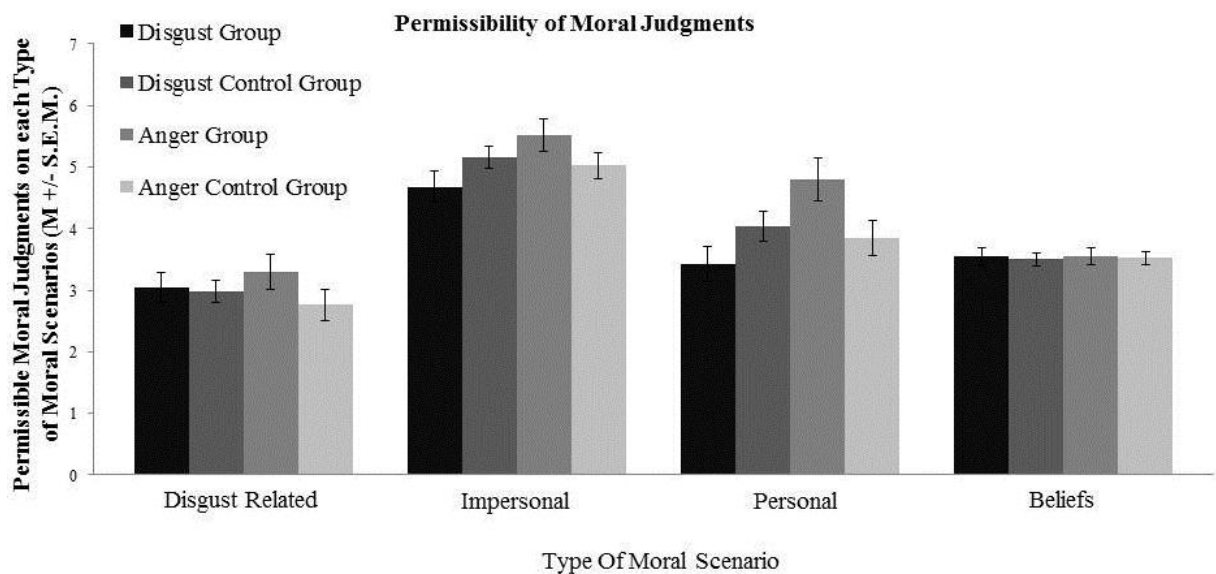
**Figure 6** Mean (+/- SEM) of affirmative judgments of moral permissibility summed over all forty moral judgments, in response to the question “Is it morally permissible for the protagonist to do x,”.

However, as expected, we observed a significant interaction Group\*Emotion and a significant main effect of Type. Furthermore, the interaction Group\*Emotion\*Type revealed a tendency to significance at the 10% level (see Table 3). Figure 6 shows an overview of the moral permissibility judgments averaged by group.

	Selected Sample			Full Sample		
ANOVA results	F-value (df)	P Value	$\eta_p^2$	F-value (df)	P Value	$\eta_p^2$
Group	$F < 1$	-	-	$F(1,282) = 2.38$	0.124	0.080
Emotion	$F(1,228) = 2.95$	0.085	0.130	$F(1,282) = 3.983$	0.047	0.140
Type	$F(3,684) = 70.428$	$< 0.001$	0.236	$F(3,846) = 98.269$	$< 0.001$	0.258
Group*Emotion	$F(1,228) = 6.989$	0.009	0.030	$F(1,282) = 2.199$	0.139	0.008
Group*Type	$F < 1$	-	-	$F < 1$	-	-
Emotion*Type	$F(3,684) = 1.756$	0.164	0.080	$F(3,846) = 1.108$	0.340	0.040
Group*Emotion*Type	$F(3,684) = 2.45$	0.073	0.011	$F(3,846) = 2.218$	0.095	0.008

**Figure 7** Mean (+/- SEM) of affirmative judgments of moral permissibility summed over all forty moral judgments, in response to the question “Is it morally permissible for the protagonist to do x,”.

Figure 7 shows the judgments of moral permissibility averaged by group for each type of moral scenario. Finally, an additional planned comparison revealed that emotions’ effect was not stronger for the personal than for the impersonal moral scenarios (see Table 4, fourth row).



**Figure 7** Mean (+/- SEM) of affirmative judgments of moral permissibility separated by the four types of moral scenarios, with 10 judgments each.



## 2.5. Discussion

The main aim of this study was to investigate the effect of different emotions on different types of moral judgments. More specifically, we predicted that two types of negative emotions differing in the motivational tendency they elicit - approach vs. withdrawal - would have differential effects on different types of moral judgments. The induction of an approach emotion such as anger was expected to increase judgments of moral permissibility, while the induction of a withdrawal emotion such as disgust was expected to decrease them. Indeed, we expected the induction of an approach emotion to augment the participant's predisposition to endorse the action described in the scenarios (e.g., pushing a person off the bridge or pushing a lever) and that they would therefore be inclined to judge such an action as morally permissible. On the contrary, the induction of a withdrawal emotion was expected to reduce the participant's willingness to endorse the described action, resulting in judgment of such actions as not permissible.

Furthermore, we predicted that the judgments of moral permissibility would not only be influenced by the type of emotion induced, but also by the type of scenario for which moral judgments were required. We thus utilized four types of frequently used moral scenarios, namely, disgust-related, impersonal, personal, and beliefs scenarios. Based on the existing evidence, we expected emotions to predominantly influence the disgust-related (Schnall et al., 2008a; Wheatley & Haidt, 2005) and personal moral scenarios (Greene et al., 2001, 2004; Koenigs et al., 2007; Valdesolo & DeSteno, 2006; Young & Koenigs, 2007), whereas a weaker effect was expected on the impersonal and beliefs moral scenarios.

The results of the manipulation checks confirmed that a) the emotion induction procedures successfully induced the intended emotions in the experimental groups (see Figure 5; note also that even though our selection criteria were tailored to ensure the general trend of these findings, the analysis of the data of the non-selected sample delivered similar results for the manipulation checks); b) a neutral emotional state predominated in the control groups; c) no relevant differences were found between the experimental groups, neither concerning the decay of emotions over time nor the participants' beliefs about the effect that emotions might have had on their moral judgments. Taken together, the results for the manipulation checks clearly show

that our experimental manipulation successfully induced the desired emotional states in participants. The unexpected mild induction of sadness in the anger group is not problematic considering that a) anger was felt significantly more strongly than sadness, suggesting that the main group effects on moral judgments are attributable to the presence of anger and that b) the expected effects of sadness - a withdrawal emotion - on moral judgments might have counteracted the effects exerted by anger. Thus, the size of the effects of anger on moral judgments in the present study might have been underestimated and might be even stronger in the absence of concomitantly induced sadness (which is a withdrawal-related emotion). Finally, as we were primarily interested in the effects of induced basic emotions, we did not control for the induction of other more complex emotions such as guilt or shame, which might have resulted from the anger induction procedure.

The significant interaction of group and emotion confirmed our first hypothesis of an effect of emotion induction on judgments of the permissibility of actions in different types of moral scenarios. More specifically, the tailored comparison of overall judgments of permissibility in the anger versus the disgust conditions confirmed that the former indeed resulted in more judgments of permissibility than the latter. These findings therefore support the hypothesis that motivational tendency is a crucial feature in determining how emotions affect moral judgments. Thereby, these results directly address the main objective motivating our study, which was to uncover the mechanisms by which emotions exert their influence on moral judgments.

Furthermore, our results also support the prediction that the influence of emotions also depends on the types of moral scenarios in which moral judgments were requested. However, not all of our predictions were confirmed. In one sense, the results confirm that emotion effects are particularly strong for moral judgments in personal moral scenarios, and basically absent for those expressed in belief moral scenarios. Contrary to our predictions, however, emotions influenced moral judgments in the impersonal moral scenarios, and not in the disgust-related ones. The latter observation was particularly surprising, as it stands against previously reported findings (Schnall et al., 2008a; Wheatley & Haidt, 2005) suggesting that the emotion of disgust plays a crucial role in moral judgment.

As mentioned, we also did not observe stronger effects of emotions on personal compared to impersonal scenarios (see also Table 4), indicating that the experimental induction of emotions also affects impersonal moral judgments and that emotion induction effects for them are similar in size to those observed for the personal scenarios. This might seem at odds with theoretical arguments in the field of moral psychology suggesting that moral judgments triggered by impersonal scenarios are less likely to entail a spontaneous emotional response than personal scenarios (Greene et al., 2001, 2004; Koenigs et al., 2007; Valdesolo & DeSteno, 2006; Young & Koenigs, 2007). The present results, therefore, point to two possible roles of emotions when judging these two similar types of moral scenarios. While previous research (Greene et al., 2001, 2004) suggests that personal scenarios *spontaneously* recruit affective processes more extensively, experimentally inducing emotions seems to affect judgments in both types of scenarios in a similar way. As a possible mechanism, we suggest that induced emotions possibly override the tendency of personal scenarios to spontaneously elicit affective states.

The specific effects of emotion induction on impersonal and personal but not disgust and belief moral scenarios observed in our data might be accounted for in several ways. A first explanation could be the fact that when evaluating the personal and impersonal moral scenarios, the subject is asked to evaluate whether it is permissible for him/her to perform the described action from a first-person perspective (e.g., is it permissible for *you* to push a man off the bridge?), while in the disgust-related and belief scenarios the action judged is performed by a third party (e.g. is it permissible for A to bribe B?). Although plausible, this explanation is not totally convincing, as previous findings document that emotions influence moral judgments in both cases, i.e., when the judged action is performed by a third party (Schnall et al., 2008a; Wheatley & Haidt, 2006), or when it is performed in first person (Valdesolo & DeSteno, 2006).

In line with Greene et al. (2009) and with our assumption that the impact of emotions on moral judgments may depend on the motivational action tendency of emotions, an alternative explanation for the unexpected effects may be that induced emotions particularly influence moral scenarios entailing a strong *action demand*. In fact, while the personal and impersonal scenarios require one to imagine performing an effortful action, such as pushing a man or a lever, the disgust related and belief scenarios do not require such action imagery. Future studies will have to provide empirical evidence to clarify which features are most likely to drive the observed

effects of emotions induction on moral judgments. Additional features such as differences in syntactic or text complexity, response times, familiarity, emotional salience, degree of conflict of each moral scenario will have to be controlled by such investigations as well.

The absence of emotion induction effects on beliefs paradigms is in line with our predictions and with previous findings reporting that the brain areas recruited to make this type of moral judgment are usually associated with detecting multi-sensory integration incongruence (e.g. visual and proprioceptive signals from the body), mentalizing, and re-orienting attention from self- to other-related processing (Decety & Lamm 2007). Furthermore, previous studies investigating moral judgments given during belief scenarios did not find an increase in activation of brain areas directly linked with affective processing during moral judgments (Greene et al., 2001; Moll et al., 2002) or with affective brain networks known to support emotions and empathy (Lamm and Singer, 2010; Singer, 2006; Singer & Lamm, 2009), reported by other studies. However, the lack of an influence of disgust induction on moral judgments for the disgust-related scenarios is surprising given the results of previous studies (Schnall et al., 2008a; Wheatley & Haidt, 2005). As possible reasons accounting for this discrepancy between current and previous findings, we can exclude an inefficiency of the chosen disgust induction procedures in Experiments 1a and 1b. Our analyses clearly confirmed that: a) the disgust induction led to reliably stronger subjective reports of feelings of disgust as compared to other emotions in both experiments, as well as to a comparable self-rated emotion intensity as compared to the anger induction, b) the disgust induction did not show any reliable effect on disgust-related moral judgments even for the participants selected on the basis of high emotion induction ratings (see Figure 5), and c) the disgust induction had considerable effects on judgments in two other types of moral scenarios for the same participants (the personal and impersonal scenarios) in both Experiments 1a and 1b.

However, the different response format used in our and in previous studies might explain the discrepancy in results. While participants in the two previous studies showing disgust-related effects (Schnall et al., 2008a; Wheatley & Haidt, 2006) were asked to quantify the wrongness of a moral violation on a scale from zero to one hundred, participants in our experiments were asked to evaluate whether the situation described in the scenario is morally acceptable or not. This may have resulted in a floor effect in the control group, leaving no space to observe a

possible effect of disgust on moral judgments. In other words, the reference baseline of judgments of permissibility given by the participants kept in a neutral emotional state in our study might already have been too low to allow for observing an effect of disgust in the disgust-related scenarios. Another possible reason for the different findings may be related to cross-cultural differences. The disgust-related scenarios used in our and previous studies were designed to represent stereotypical violations of US American moral norms and they may have had different effects on the Swiss-German sample investigated in our study.

In summary, our study allows us to draw some important conclusions on the relationship between emotions and moral judgments. First, our results show that experimentally induced emotions play a direct causal role in determining moral judgments, contradicting arguments in moral psychology that emotions are not involved in moral judgments. Second and most importantly, our findings contribute to a better understanding of the mechanisms by which emotions influence moral judgments, showing that action motivational tendencies predict how moral judgments will be influenced by emotions: while anger as an approach emotion increased judgments of moral permissibility, the withdrawal emotion of disgust decreased them. We propose that this effect was modulated by the differences in the action demands implied by moral scenarios. As both impersonal and personal moral scenarios involve stronger action demands compared to disgust-related and beliefs scenarios, the induction of emotions had a stronger influence on the former two.

In conclusion, the present work demonstrates that experimentally induced emotions causally influence moral judgments and that this influence critically depends on the type of emotion induced and the type of moral scenario evaluated. We believe that the role played by more complex moral emotions such as contempt or indignation also strongly depends on the elements involved in the moral scenarios judged and on the motivational tendencies these complex emotions induce. Our data are compatible with the view that moral judgments result from a combination of both emotional intuitions and reasoning, and that the relevance of each of these mechanisms depends on the moral scenario being evaluated. To get a better understanding of the mechanisms by which emotions influence moral judgments, future studies will have to determine how factors such as valence, motivational tendency of emotions, and action demands of the decisions collaborate in influencing moral judgments. For this, a better taxonomy as well

as better standards of normative empirical data on the psycholinguistic, emotional and other variables characterizing the moral scenarios used in research is needed, allowing for a priori classification of the cognitive, affective, and action-related components needed to make judgments about different types of moral scenarios (see also Knutson et al., 2010). Finally, the present study focused only on the role of different types of emotions on moral judgments but ignored moral actions. The observed effects of emotions on moral decision making and behavior may be even stronger when participants are actually required to perform moral actions, and not only to judge their permissibility.

### **3. The Causal Role of the LPFC in Social Norm Compliance.**

(Adapted from Ruff, C., Ugazio, G., Schlaepfer, A. and Fehr, E. (Submitted.)).

#### **3.1. Abstract**

Social coexistence crucially depends on compliance of social norms and the punishment threat associated with norm violation. Previous studies suggest that brain activity in the lateral prefrontal cortex (LPFC) is processing the punishment threat attached to norm violation, therefore correlating with social norm compliance. In this study we test the causal functional involvement of the LPFC in behavioral regulation in the context of social punishment threats for norm-violating behavior. We hypothesized that, compared to a neutral control condition, participants with enhanced LPFC functionality would display a more norm compliant behavior, while an increase in norm-defecting behavior was expected by participants with impaired LPFC functionality. To assess the causal functional role of LPFC for norm-compliance participants received anodal, cathodal, or sham transcranial direct current stimulation (tDCS) while making decisions on how to distribute a certain amount of money between themselves and a second person, in one treatment where proposers may be punished for social norm violations and a control treatment where such punishment was not allowed. The data obtained in this study show for the very first time that LPFC function is causally necessary for social norm behavior as experimentally decreased LPFC activity determined an increase in norm-defecting behavior, while an increase in LPFC activity resulted in an increase in norm-compliant behavior. We conclude that it is possible to manipulate social behavior through targeted brain stimulation and that the LPFC plays a crucial causal role in determining human likelihood to comply with social norms.

#### **3.2. Introduction**

Social interactions are often regulated by social norms, defined as widely shared beliefs about what constitutes an acceptable standard behavior in a given situation (Elster, 1989). Despite the prevailing acceptance of such norms, their maintenance is at constant risk as some individuals are prone to violating them, given that most of the times the appropriate behavior requires a person to sacrifice a personal gain in order to promote the society's interest. Such

defectors can have a strong destabilizing effect on social norms as most of the people usually comply with norms only conditional on others' compliance (Fischbacher et al., 2001). To preserve social norms from decaying, some form of punishment threat is typically required, be it through the legal enforcement system or more informal peer punishment within a small group of individuals (Fehr and Gächter, 2002). Among these, informal peer punishment held a prominent role in the evolution of human sociality (Sober and Wilson 1998) leading some to assume that the human brain may have developed specific neural mechanisms processing punishment threats in the context of social norm enforcement.

Evidence for such a conjecture was recently given in a neuroimaging study (Spitzer et al., 2007), measuring brain activity in an economic paradigm specifically constructed to test the effect of punishment threat on people's compliance with the fairness norm while allocating money. Results show that the prospect of punishment associated with norm-violating behavior led participants to strongly adjust their decisions towards norm-compliant choices, and that these norm-compliant behavioral adjustments were strongly related to right-lateralized activation of the dorso-lateral prefrontal cortex (DLPFC) and orbito-frontal cortex (OFC). While this study clearly established an involvement of lateralized prefrontal cortex (LPFC) in the guidance of norm-compliant behavior, due to the well-known fact that neuroimaging studies can only provide correlative data on the relation between brain functions and behavior (Driver et al., 2009), it cannot give any information on whether activity in this brain structure is indeed causally involved in the generation of norm-compliant behavior. Evidently, establishing which brain networks are causally determining the willingness of a person to comply with the norms is of critical importance as it would revolutionize our understanding of social and antisocial behaviors.

A well-established method to determine the causal relationship between brain activity and behavior can be achieved in healthy participants with brain stimulation methods, such as transcranial direct current stimulation (tDCS). In this study we assessed the causal functional involvement of the LPFC in behavioral regulation in the context of social punishment threats for norm-defecting behavior. More specifically, the causal functional role of the LPFC for participants' likelihood to comply with the fairness norm was assessed via an economic decision



making paradigm similar to the one introduced by Spitzer and colleagues (Spitzer et al. 2007) while receiving anodal, cathodal or sham tDCS over the LPFC.

The decision paradigm used in the present study entailed two players, A and B, who anonymously interacted with each other, knowing that they were facing either a human player, in Experiment 1, or that player B was played by a Computer, in Experiment 2. Player A received an endowment of 100 money units (MUs) which could be freely allocated between herself and player B. In the punishment condition player B could punish A after being informed of the latter's decision (resembling an ultimatum game, Andreoni et al. 2003), while in the non-punishment condition (resembling a dictator game, Kahneman et al., 1986) player B was a passive recipient (for further details see the Experimental Procedures, and Spitzer et al., 2007). Previous behavioral evidence suggests that victims of the fairness norm violation are motivated to punish the norm defectors (Fehr and Fischbacher, 2004; Fehr and Gächter, 2002; Güth et al., 1982), therefore the latter face a concrete punishment threat when player B is allowed to punish. Furthermore, it has also been shown that the threat resulting from the presence of potential punishment is sufficient to induce proposers to respect the fairness norm (Fehr and Gächter, 2002, Spitzer et al., 2007).

In the neuroscientific literature several studies reported strong correlations between brain activity in the prefrontal brain regions and the inhibition of prepotent responses (Aron et al., 2004; Miller and Cohen, 2001; Sanfey et al., 2003) as well as in the evaluation of punishing stimuli (Kringelbach, 2005; O'Doherty et al., 2001). In the previously mentioned research, Spitzer and colleagues (2007) have shown that prefrontal regions are activated by the presence of a punishment threat in the mentioned economic decision making task. Additionally, a strong positive correlation between brain activity in the DLPFC and the increase in norm compliance it has been reported (Spitzer et al., 2007). Taken together, these data suggest that the mentioned prefrontal regions are involved in norm compliant behaviors by constraining the individual selfish impulses.

Further evidence suggesting that the LPFC is involved in controlling impulses is provided by studies on intertemporal choice (Figner et al. 2010). The intertemporal choice task typically requires participants to decide between receiving a smaller amount of money soon and a larger

amount of money after a longer period of time. Studies with such standard tasks usually vary two variables, a) the difference between the smaller and the larger amount and b) the dimension of the time interval that one has to wait in order to get the larger amount of money (Green and Myerson, 2004). Several studies have shown that humans and animals have a strong preference (or impulse) towards rewards delivered sooner, in particular if a reward may be obtained immediately (Green et al., 2004). In the mentioned study, Figner and colleagues (2010) used transcranial magnetic stimulation (TMS, a brain stimulation method similar to tDCS) to assess the causal role of the LPFC for impulse control. To test this, they measured intertemporal choices of participants with either normal brain activity or impaired LPFC brain activity. The evidence provided suggests that the LPFC has a causal role in controlling impulses, as participants with reduced LPFC activity chose significantly more often the sooner choice than those with unaltered LPFC activity.

Based on this concise literature review it seems that the LPFC has a critical role in controlling impulses, prompting people to make more “cold-blood” decisions instead of rushing to take the option appearing at first more attractive. We therefore hypothesized that the neural networks embedded in the LPFC may have a causal functional role in determining individuals’ readiness to comply with norms in the presence of a punishment threat. In this case in fact the capacity of controlling the greedy impulse is required in order to avoid taking a too large share large, which if considered unfair by the one’s counterpart would lead to punishment. We therefore predict that in the presence of punishment a) an enhanced brain activity in this region would result in a more norm compliant behavior, and b) suppression of brain activity in this region would result in a more selfish, norm-violating, behavior. More in detail, we expected participants with greater LPFC activity to both share a greater amount of money during the punishment condition and to display a greater preference update (measured as the difference in money transferred during trials where punishment was allowed minus those trials not entailing any punishment) deriving by the punishment threat’s presence or absence, compared to those participants who received no stimulation and those in whom LPFC activity was disrupted. On the contrary, these latter are expected to share both a smaller amount of money in the punishment condition and display a smaller update of preference from the situation where punishment was

allowed to the one where punishment was not allowed, compared to the non-stimulated participants and those with enhanced LPFC activity.

To test whether the LPFC causal functional role is specific for social punishment threats, and not for other variables such as risk, in a control experiment (see Experiment 2 below) participants received anodal cathodal and sham stimulation while performing in the very same paradigm but knowing that they were facing a computer and not a human opponent (see also Spitzer et al., 2007). In this case we hypothesized no effect from any of the two stimulations over the LPFC on participants' choices.

### **3.3.Methods**

This study was approved by the local ethics committee and participants were fully informed about the experimental procedure beforehand in a consent form which they had to read and sign before the beginning of the study.

#### **Experiment 1**

##### *Participants*

To exclude any gender-dependent variance only females participated in our study. Seventy-seven undergraduate students of the University of Zurich (age,  $M = 22$ ,  $SEM = 0.358$ , Range = min 18, max 32) took part in experiment 1 in exchange for money (25 Swiss Francs (CHF) per hour plus what they earned in the game). Participants were randomly assigned to one of three groups. Groups differed with respect to the type of tDCS stimulation (see below for details) they received: anodal ( $n = 25$ ), sham ( $n = 25$ ), and cathodal ( $n = 27$ ). Twelve people, five from the anodal and sham group and two from the cathodal group, were excluded from the analysis as manipulation checks and their behavior during the task revealed that they did not understand the task by failing to report their answers within the allowed response time (10s) and the other 5 did not report any answer in several of the 24 trials). An additional participant in the cathodal group was excluded for moving the tDCS electrode resulting in the failure of the stimulation while performing the task. Therefore, after exclusion, 63 participants (19 assigned to the anodal stimulation group, 20 to the sham group, and 24 assigned to the cathodal group) were included in the analyses.

## *Procedure*

The experimental paradigm employed here closely follows the procedure previously used in Spitzer et al. 2007: two players (A and B) are endowed with 100 money units (MUs), corresponding to 1 CHF (or 1MU = 0.01 CHF) and have to divide these among themselves. Both A and B receive 25 MUs extra on every trial, for reasons of fairness and to make social punishment possible. Player A, or “the proposer”, can always decide how the 100 MUs will be allocated between A and B. However, only in the baseline condition of the paradigm this decision is implemented straight away. In the punishment condition, Player B, or “the responder”, can spend part or all of the extra 25 MUs to punish the proposer, with the rule that each MU invested by B in the punishment leads to the removal of 5 MUs from A’s gain. Both players are cued whether the upcoming trial is in the baseline or punishment condition; this means that B is always aware whether she will possess the option for social punishment if he considers A’s choice as violating the fairness norm, and A is always aware of this social punishment threat. Given that the aim of this experiment was to test on the effects of tDCS on Players’ A norm compliance, players’ B were not present during the stimulation sessions, but their answers were hard-coded in the program used in the experiment. Participants in these sessions were therefore told that they were playing with a human counterpart, but they did not know he was not part of their experimental session.

## *Manipulation Checks*

To control for an eventual effect of tDCS on the expectations participants had on the influence of the stimulation on their behaviour, at the end of the study but while the stimulation was still running, participants reported on a four-point scale (ranging from “Gar nicht”, “not at all” in German, to “Sehr”, “a lot” in German) whether a) how strongly would Player B punish Player A for an offer of respectively 20, 40 or 60 per cent of her endowment, b) how angry they believed Player B would be when he received an offer of respectively 20, 40 or 60 per cent of Player’s A endowment, and c) how fair they believed is Player’s A offer of respectively 20, 40 or 60 per cent of her endowment to Player B.

## *tDCS Stimulation*

During the experiment, we applied tDCS over participant's LPFC. This technique allows for modulation of regional neural excitability by means of applications of weak currents. In short, neural activity (i.e., an action potential) is usually elicited if the membrane potential – usually is -65 and -35 mV - at rest – is lowered to about -30mV (Bear et al., 2001), via driving inputs through other neurons. Applying weak currents over a cortical area via tDCS can increase or decrease the resting membrane potential, depending on the position and polarity (anodal or cathodal) of the electrode. Thus, tDCS can lead to an increase or decrease of the excitability and spontaneous activity in the neural tissue under the electrode. In the present study, we applied anodal, cathodal, or sham tDCS over the right LPFC, using a standard CE-approved stimulator and a set of standard 35cm<sup>2</sup> electrodes (both produced by the Neuroconn company Ltd., Whitland, UK). The stimulation point was defined using the coordinates reported by Spitzer et al. (2007) corresponding to the highest registered brain activity in the LPFC and localized in each participants using Brainsight 2.0. The reference electrode for anodal and cathodal stimulation was positioned over the vertex. In line with established procedures (see Iyer et al., 2005), we stimulated for 20 minutes with 1 mA for the anodal and the cathodal condition, or for 20 seconds in case of the sham placebo stimulation. This latter condition feels identical to the anodal and cathodal condition, but has no measurable effect on neural excitability.

## **Experiment 2**

### *Participants*

Sixty-four female undergraduate students of the University of Zurich participated in exchange for money (see the corresponding section for Experiment 1) in Experiment 2 (age,  $M = 22$ ,  $SEM = 0.339$ , Range = min 18, max 29), again divided in three groups: anodal ( $n = 23$ ), sham ( $n = 21$ ), and cathodal ( $n = 20$ ). The same criterion used in Experiment 1 to exclude participants who did not comply with the instructions was also applied in Experiment 2, resulting in the exclusion of 5 participants (2 from the anodal and three from the sham groups) who did not report any answer in some of the 24 trials. In total we therefore included in the final analyses fifty-nine participants, 21 in the anodal, 18 in the sham and 20 in the cathodal groups respectively.

## *Procedure*

Experimental procedures were the same as in Experiment 1 with the exception that this time the role of Player B was simulated by a computer. Participants were explicitly informed about this at the beginning of the experimental sessions, and therefore they knew they were not playing with a human counterpart.

## *Statistical Analysis*

The data from the experiments consists of a panel with 123 individuals ( $i$ ) that play for 24 periods ( $t$ ) each. We perform a regression analysis in which we aim at explaining the determinants of the amount that participants transfer to their counterparts. Therefore the dependent variable is TRANSFER, i.e. the money allocated by each participant in each round, which varies from 0 to 100. The independent variables TDCS (3 factor levels: Anodal, Sham and Cathodal), SOCIAL (2 factor levels: Social and No-Social), and PUNISHMENT (2 factor levels: Punishment and No-Punishment), were constructed based on the brain-stimulation type received by participants, whether the experiment was a social or a non-social one, and whether punishment was allowed or not, respectively, following the experimental manipulations previously described in the methods section. Furthermore to test whether the independent variables modulate each other, e.g. if the effect of the TDCS is stronger in a social environment when punishment is allowed, two-way interactions and a three-way interaction between these independent variables were included among the explicative variables of our regression. Additionally, among the independent variables were included those capturing the strategic and risk-propensity of participants' personality (see the methods section). Formally, the model we test is expressed in the following equation:

$$\begin{aligned} Transfer_{i,t} = & TDCS_i + Social_i + Punishment_{i,t} + TDCS_i * Social_i + TDCS_i * Punishment_{i,t} + Social_i \\ & * Punishment_{i,t} + TDCS_i * Social_i * Punishment_{i,t} + \eta_i + \nu_t + \varepsilon_{i,t} \end{aligned}$$

where the  $i$  and  $t$  subscripts indicate whether a given variable varies over individual and/or time. Note that the error component contains a time invariant component  $\eta_i$ , capturing the individual specific unobserved characteristics, and a time specific component  $\nu_t$ , capturing the effect that a specific time period may have on our dependent. Indeed, if these omitted components were to be

correlated with any of our independent variables of interest, the coefficients in the regression procedures could be biased due to endogeneity problems. Typically to rule out such endogeneity concerns one would perform a fixed effects estimation. Despite our variable of interest, i.e. TRANSFER, varies over individuals and time periods, in our case, however, it is not feasible to use a fixed effect estimation as two of the main independent variables, TDCS and SOCIAL, vary solely across individuals by experimental design (see the methods section). Consequently, given that if we were to perform a fixed effects estimation these time invariant variables would be omitted, we are required to perform a random effects regression. The main assumption made when performing this estimation is considering the time invariant individual specific error component not correlated with the independent variables under examination  $X$ , i.e.  $\text{corr}(\eta_i, X) = 0$ . Our experimental design provides several points supporting this assumption, for instance participants in this study have similar characteristics being a) all females b) of similar age, and c) students of Zurich University (see the methods section). Further and most importantly, all of them were randomly allocated to each experimental group, assigned depending on the type of TDCS received and of playing with a social or non-social counterpart, should guarantee that both  $\text{corr}(\eta_i, TDCS_i) = 0$  and  $\text{corr}(\eta_i, Social_i) = 0$ .

Provided the above details on our statistical sample, the best estimation model corresponds to a random effects generalized least square (GLS) regression. Indeed, the GLS estimation is also needed to rule out the potential heteroskedasticity of the error term which in our case includes several components, given the panel nature of the data. To control for differences in within-participant variance we cluster the participants' standard errors using the bootstrap approach (repeated 10000 times), considered that bootstrapping is the most rigorous procedure to control for the mentioned variance in a GLS model. This bootstrap clustering of standard errors was adopted in all the regressions described hereafter.

As the results revealed a significant effect of the interactions included in the first regression (see Table 6 in the results section), we performed two additional GLS regressions in which we tested specifically the effects of TDCS and PUNISHMENT, as well as their interaction, on TRANSFER, for the participants in the *social* and *non-social* groups respectively. Further, we performed two additional regressions solely within the *social* sample, one for the monetary allocations made where punishment was allowed and the other for those where

punishment was not allowed, testing the effect of TDCS on the dependent variable TRANSFER. Indeed, the results of the previous two regressions yielded a significant effect of TDCS\*PUNISHMENT in the *social* sample but not in the *non-social* one (see Table 7 in the results section).

### 3.4.Results

Table 5 below displays the summary statistics of the dependent, independent and control variables.

Variable	Obs	Mean	Std. Dev.	Min	Max
Transfer	3035	27.26524	23.73012	0	100
tDCS	3048	-0.01575	0.822885	-1	1
Punishment	3048	0.5	0.500082	0	1
Social	3048	0.496063	0.500067	0	1
Tdcs_Punishment	3048	-0.00787	0.581921	-1	1
Tdcs_Social	3048	-0.03937	0.58064	-1	1
Social_Punishment	3048	0.248032	0.431941	0	1
Tdcs_Social_Punishment	3048	-0.01969	0.411047	-1	1
Period	3048	12.5	6.923322	1	24
Age	3048	21.55906	2.969231	18	32
Income	3048	2.464567	1.078198	1	4
Socialstatus	3048	2.23622	0.917668	1	4
Peopleknown	3048	0.267717	0.798247	0	5

**Table 5:** Summary statistics of the dependent, independent, and control variables (used for robustness checks).

The results of the regression measuring whether transfer was influenced by the variables TDCS, SOCIAL, and PUNISHMENT, as well as the interactions TDCS\_PUNISHMENT, SOCIAL\_PUNISHMENT, and TDCS\_SOCIAL\_PUNISHMENT, revealed a significant effect at 0% confidence level. Further, the interaction TDCS\_SOCIAL also revealed a significant effect at 3% confidence level (see Table 6). More specifically, TDCS had a negative effect on transfer, meaning that the stronger the excitability in the LPFC the smaller the amount of money transferred by a participant. Further, SOCIAL and PUNISHMENT had a positive effect on TRANSFER indicating that the money transferred to the counterpart increased when a participant was allocating money either to another person or could be punished by the other person. However, given the significant results of all the interactions, to fully understand how these independent variables affected transfer it is necessary to investigate deeper.



Random-effects GLS regression			Number of obs		=	3035
Group variable: id			Number of groups		=	127
R-sq: within = 0.6450			Obs per group: min		=	20
between = 0.0802			avg		=	23.9
overall = 0.5689			max		=	24
corr(u_i, X) = 0 (assumed)			Wald chi2(7)		=	6367.22
			Prob > chi2		=	0.0000
(Replications based on clustering on id)						
transfer	Observed Coef.	Bootstrap Std. Err.	z	P> z	Normal-based [95% Conf. Interval]	
tdcs	-1.008729	.5167081	-1.95	0.051	-2.021459	.0039998
punishment	41.31893	.6440263	64.16	0.000	40.05666	42.5812
social	11.35904	.7240692	15.69	0.000	9.939889	12.77819
tdcs_punishment	2.003258	.7772913	2.58	0.010	.4797953	3.526721
tdcs_social	-1.733068	.8854586	-1.96	0.050	-3.468535	.0023985
social_punishment	-13.09527	.9906169	-13.22	0.000	-15.03684	-11.15369
tdcs_punishment_social	3.62774	1.214986	2.99	0.003	1.246411	6.009068
_cons	4.156555	.4267174	9.74	0.000	3.320204	4.992905
sigma_u	8.1267738					
sigma_e	13.445521					
rho	.26757441	(fraction of variance due to u_i)				

**Table 6:** GLS regression testing the effects of TDCS, PUNISHMENT, and SOCIAL, as well as their interactions TDCS\*PUNISHMENT, TDCS\*SOCIAL, SOCIAL\*PUNISHMENT and the three way interaction TDCS\*PUNISHMENT\*SOCIAL on the dependent variable TRANSFER

The second regression we performed, revealed that within the participant in the *social* sample (i.e. those allocating money to another person) a main significant effect of TDCS, PUNISHMENT, and of the interaction of these two variables TDCS\*PUNISHMENT on TRANSFER at a 0% confidence level.

Random-effects GLS regression		Number of obs = 1619				
Group variable: id		Number of groups = 68				
R-sq: within = 0.5111		Obs per group: min = 20				
between = 0.0033		avg = 23.8				
overall = 0.4181		max = 24				
corr(u_i, X) = 0 (assumed)		Wald chi2(3) = 1731.21				
		Prob > chi2 = 0.0000				
(Replications based on clustering on id)						
transfer	Observed Coef.	Bootstrap Std. Err.	z	P> z	Normal-based [95% Conf. Interval]	
tdcs	-3.181293	.6926802	-4.59	0.000	-4.538921	-1.823664
punishment	28.60308	.719341	39.76	0.000	27.1932	30.01297
tdcs_punishment	5.790991	.905493	6.40	0.000	4.016258	7.565725
_cons	14.64188	.5586766	26.21	0.000	13.54689	15.73686
sigma_u	9.3127429					
sigma_e	14.345523					
rho	.29648167	(fraction of variance due to u_i)				

**Table 7:** GLS regression testing the effects of TDCS and PUNISHMENT, as well as their interactions TDCS\*PUNISHMENT, on the dependent variable TRANSFER solely in the *social* sample

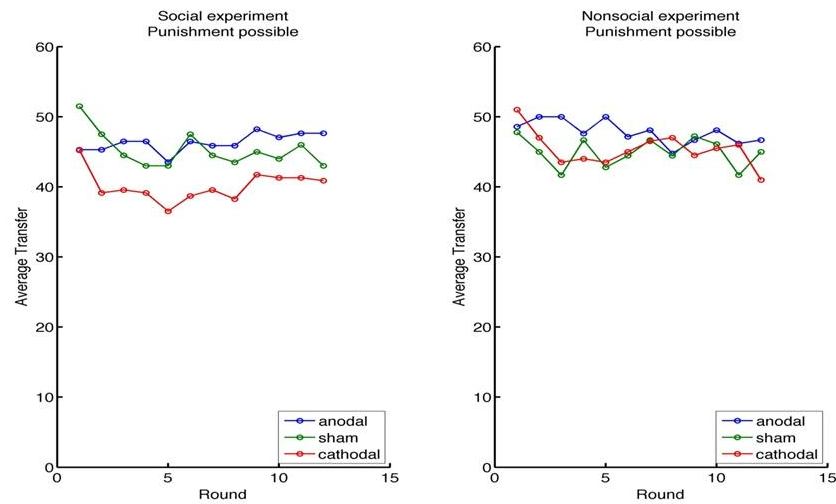
In details, TDCS had a negative effect on transfer, meaning that participants with increased excitability in the LPFC were transferring less money than the others; further PUNISHMENT had a positive effect on transfer indicating, as in the previous regression, that when participant could be punished by their counterparts, their monetary transfers increased (see Table 7 above). Again, a deeper investigation is required to fully grasp the effect of TDCS on TRANSFER as the interaction of TDCS with PUNISHMENT is significant. Beforehand, however, it is important to report the results from this very same regression ran this time on the observations from the *non-social* sample. Importantly, indeed, this analysis revealed that, as expected, PUNISHMENT had a positive and significant effect at 0% confidence level on TRANSFER, nor TDCS nor the interaction TDCS\*PUNISHMENT had a significant effect on participants' TRANSFER ( $p = 0.30$  for TDCS and  $p = 0.07$  for the interaction TDCS\*PUNISHMENT, see Table 8 below) within this sample (i.e. those allocating money to a computer). These results, hence, seem to suggest that the main effect of TDCS observed in the initial regression (see Table 6) is specific for the participants in the social sample.

Random-effects GLS regression		Number of obs	=	1595		
Group variable: id		Number of groups	=	67		
R-sq: within	= 0.7438	Obs per group: min	=	20		
between	= 0.0280	avg	=	23.8		
overall	= 0.6832	max	=	24		
corr(u_i, X) = 0 (assumed)		Wald chi2(3)	=	4478.71		
		Prob > chi2	=	0.0000		
(Replications based on clustering on id)						
transfer	Observed Coef.	Bootstrap Std. Err.	z	P> z	Normal-based [95% Conf. Interval]	
tdcs	-.5198364	.5079857	-1.02	0.306	-1.51547	.4757972
punishment	41.00799	.6304789	65.04	0.000	39.77227	42.2437
tdcs_punishment	1.33917	.7551199	1.77	0.076	-.1408378	2.819178
_cons	4.319529	.4209156	10.26	0.000	3.494549	5.144508
sigma_u	6.879662					
sigma_e	12.322083					
rho	.23764272	(fraction of variance due to u_i)				

**Table 8:** GLS regression testing the effects of TDCS and PUNISHMENT, as well as their interactions TDCS\*PUNISHMENT, on the dependent variable TRANSFER solely in the *non-social* sample

Within the social group, we successively performed to additional analyses to disentangle how the stimulation received by participants affected their monetary allocations in the trials in

which they could be punished and in those where they could not be punished. In both cases results revealed a significant effect of TDCS on TRANSFER with a confidence level of 1%.



**Figure 8:** Average monetary transfer in the 12 rounds where punishment was allowed. On the left pane the results from the social sample are displayed, while on the right pane are the results from the non-social sample. The blu line denotes transfers from participants in the anodal group, the green line for the sham group, and the red line for the cathodal group (Color and side used in all the figures of this study are the same and will not be reported in the following legends).

Most importantly, however TDCS had opposite effects in these two situations: when punishment was allowed, TDCS had a *positive* effect indicating that a higher LPFC excitability, induced via anodal stimulation, resulted in a higher amount of money transferred from Player A to Player B (see Table 9 hereafter and Figure 8 above).

Random-effects GLS regression					Number of obs	=	899
Group variable: id					Number of groups	=	68
R-sq: within = 0.0000					Obs per group: min	=	12
between = 0.0127					avg	=	13.2
overall = 0.0013					max	=	23
corr(u_i, X) = 0 (assumed)					Wald chi2(1)	=	12.04
					Prob > chi2	=	0.0005
(Replications based on clustering on id)							
transfer	Observed Coef.	Bootstrap Std. Err.	z	P> z	Normal-based [95% Conf. Interval]		
tdcs	1.549176	.4464158	3.47	0.001	.6742169	2.424134	
_cons	41.1196	.368888	111.47	0.000	40.3966	41.84261	
sigma_u	11.111885						
sigma_e	13.071271						
rho	.41950569	(fraction of variance due to u_i)					

**Table 9:** TDCS effects on the dependent variable TRANSFER in the *social* sample and punishment rounds

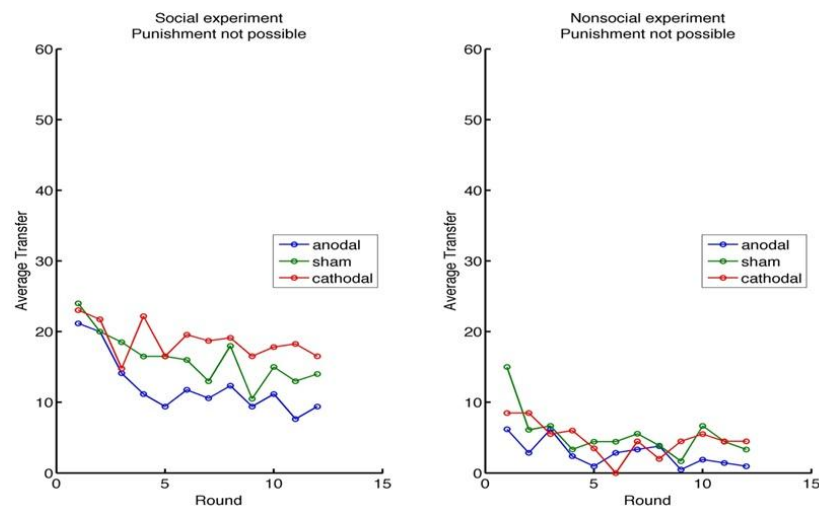
Conversely, in the situations in which punishment was not allowed, TDCS had a *negative* effect on TRANSFER, meaning that a higher LPFC excitability resulted in lower amounts of money transferred (see Table 10 and Figure 9 below).

Random-effects GLS regression		Number of obs	=	899
Group variable: id		Number of groups	=	68
R-sq: within = 0.0000		Obs per group: min	=	12
between = 0.0106		avg	=	13.2
overall = 0.0023		max	=	23
corr(u_i, X) = 0 (assumed)		Wald chi2(1)	=	10.16
		Prob > chi2	=	0.0014
(Replications based on clustering on id)				

transfer	Observed Coef.	Bootstrap Std. Err.	z	P> z	Normal-based [95% Conf. Interval]	
tdcs	-1.887812	.5922086	-3.19	0.001	-3.04852	-.7271045
_cons	16.85826	.4799431	35.13	0.000	15.91759	17.79893
sigma_u	14.723312					
sigma_e	15.386116					
rho	.4779975	(fraction of variance due to u_i)				

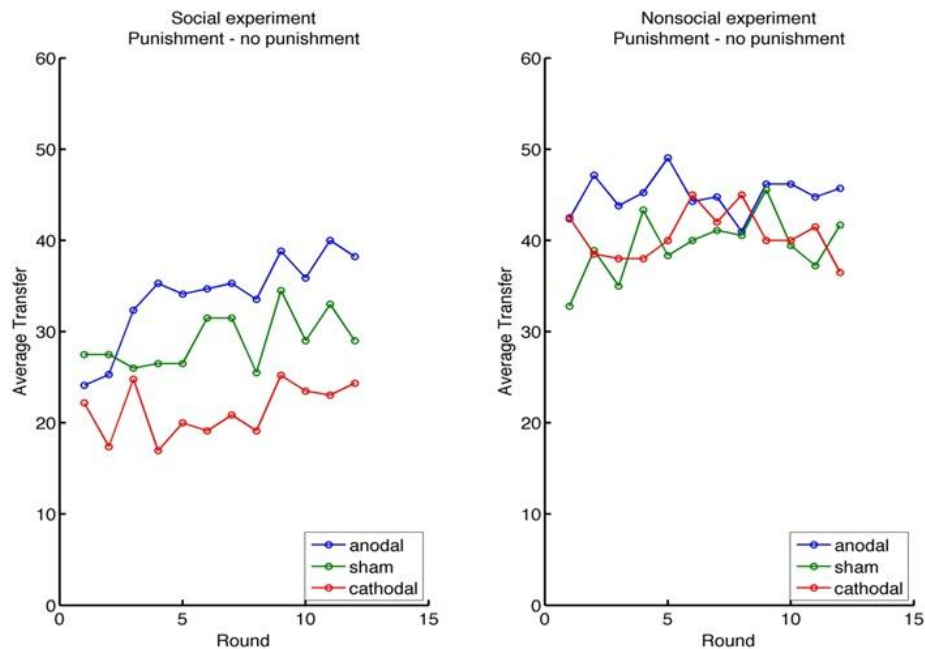
**Table 10:** TDCS effects on the dependent variable TRANSFER in the *social* sample and no-punishment rounds

Taken together the analyses performed revealed a clear pattern indicating that participants with higher LPFC excitability were those more capable of updating their allocation strategies between the two environmental conditions: when a social norm was enforced they increased the amount of monetary transfer to their counterparts in order to meet the expectations from these latter given the saliency of the norm;



**Figure 9:** Average monetary transfer in the 12 rounds where punishment was not allowed.

on contrast, when no social norm was explicitly enforced, these participants were those who kept a higher amount of money for themselves, although, on average, still transferring about 10% of the transferrable money to their counterparts. Figure 10 presents a visual representation of this norm-dependent allocation strategies updates, illustrating the difference between the money transferred by participants in the trials where they could be punish and the amount transferred in the trials where they could not be punished.



**Figure 10:** Average difference of monetary transfers (transfers while punishment is allowed minus transfer where punishment is not allowed) over the 12 rounds.

Furthermore, comparing the right with the left panels of figures 8, 9 and 10 another critical aspect of the statistical results previously reported can be clearly noted: TDCS specifically affects only monetary allocations performed in a *social context*.

### 3.5.Discussion

The main aim of the present study was to establish if the neural networks embedded in the LPFC have a direct functional causal role in determining norm-compliant behavior. In order to assess this we applied anodal, cathodal and sham tDCS stimulation over the LPFC to participants who performed a specifically designed economic paradigm (see the methods section and Spitzer et al. 2007). We hypothesized that increased LPFC neural excitability would result in

increased norm compliant behavior, while decreased LPFC neural excitability would result in the opposite effect, i.e. a decreased norm compliant behavior compared to normal LPFC excitability interacting with another person.

Furthermore we also aimed at disclosing whether the causal involvement of the LPFC on norm-compliant behavior is specific for social environments. We therefore ran two experiments: in the first participants had to allocate money between themselves and a human counterpart, while in the second one participants allocated money between themselves and a counterpart simulated by a computer program. The punishment probabilities were kept equal in both experiments, guaranteeing that any difference in the effects on behavior of the stimulation over the LPFC would be solely due to the qualitative difference of participants' counterparts. We predicted a causal involvement of the LPFC solely on behavior in the social domain.

First and most importantly, the results obtained in the present study revealed a remarkable causal involvement in motivating norm-compliant behavior for the neural networks embedded in the LPFC. The evidence provided in fact shows that when the fairness social norm was informed by means of threat punishment, participants with an increased LPFC excitability, obtained by means of anodal tDCS, were those whose behavior was most in line with the one prescribed by the norm (see figure 1). In other terms, they were those who transferred more money in this condition compared to participants with normal or reduced LPFC excitability.

Considering the literature discussed in the introduction (Figner et al. 2010; Spitzer et al., 2007 ; Aron et al., 2004; Miller and Cohen, 2001; Sanfey et al., 2003), this behavioral effect is most likely resulting from an increased capacity or regulating selfish impulses by the participants with higher LPFC excitability. This conjecture is further strengthened observing the behavior of participants with reduced LPFC excitability, i.e. those who received cathodal tDCS: they were in fact those who allocated least money in the trials where the fairness norm was enforced. In line with Figner et al. (2010), where participants with reduced LPFC activity revealed a more impulsive behavior when performing in the intertemporal choice task, participants in our study were those less capable of regulating the selfish impulse of keeping a larger share of money.

Considering the allocations made by participants in the trials where punishment was not allowed (see figure 2), our results revealed a completely reversed pattern: an increase of LPFC

excitability lead people to transfer a smaller amount of money to their counterparts, compared to those with normal or reduced excitability. These last were those who actually transferred most in this situation. This result might seem to be at odds with the interpretation that the LPFC is regulating norm-compliant behavior, or regulating impulses. However, this is not necessarily true. Firstly, the absence of punishment gives the impression to the proposers there is no specific norm enforced to regulate the monetary allocation. In the absence of a norm proposers have no other reason to transfer money to their counterparts than their own internal motivations, ranging from building a good reputation (Kocha & Normann, 2008), e.g. not appearing selfish, to warm glow (Andreoni, 1990), i.e. being kind to the other makes one feel good. Given the absence of a norm, therefore, it is not problematic that people with increased LPFC excitability are transferring less money. Rather, from a strict old-fashioned homo-economicus point of view (von Neumann and Morgenstern, 1947), it could be argued that in this type of situations the only norm present indicates to keep everything. Thus, one may even concede that participants with increased LPFC excitability are sharing less money in this condition because they are complying with a homo-economicus style norm and that therefore also in this case an increased LPFC results in an increased norm-compliant behavior.

Further, the impulse control role we attribute to the LPFC is not actually at odds with the observed reduction in the amount of money transferred. In fact, giving away money in this situation can be considered an impulse driven action resulting from the aversion of leaving their counterparts empty-handed. Considering the two motivations given as examples above, being more self-controlled would lead one to realize that giving money to try build a good self-reputation is futile being the interactions fully anonymous. Similarly, increased self-control may be leading to realize that the reward derived from the warm-glow effect of giving something to the other is actually smaller than the reward he derives from keeping a larger share of the pie for himself.

In sum, the results described in the present work suggest that the LPFC has a causal role in norm-compliant behavior, achieved by increasing the self-control abilities of a person. In other words an increased excitability in the LPFC resulted in a better capacity in participants to behave strategically quickly adapting their behavior to the accepted standards given the particular situation. In the present study we are therefore providing the previously missing necessary

evidence to corroborate the speculative hypothesis put forward (Spitzer et al. 2007) regarding a casual involvement of the LPFC in norm compliant behavior

Furthermore, our results demonstrate that LPFC has causal role in determining norm-compliant behavior solely in a social context: our analyses revealed that tDCS effects were significantly affecting monetary transfers in participants of the social group, but such effect disappeared in participants within the non-social group. This remarkable result therefore suggests that the LPFC is selectively involved in behaviors happening in a social environment. Furthermore, the direct comparison of experiments 1 and 2 allows us to claim a specific role of LPFC for norm-compliant behavior, ruling out many alternative potential confounding explanations of the observed behavioral effects. For instance, one could have claimed that the LPFC is affecting participants' risk propensities, but if this was the case a similar effect of tDCS should have been found in both experiments 1 and 2 and this was not revealed by the data presented here.

Further, consistently with previous literature (Fehr and Fischbacher, 2004; Fehr and Gächter, 2002) our results revealed a critical role of punishment in the allocating behavior, regardless of the stimulation participants received or the qualitative type of the sample (social or not). All the analyses in which the effect of punishment was tested revealed the well-established pattern of punishment systematically increasing the amount of money a proposer allocates to his counterpart. Replicating this well established effect contributes to strengthen our confidence in the other effects found in the present study.

In conclusion, our paper has for the very first time established a direct causal link between functional brain activity, in the present case located in the LPFC, and norm regulated behavior. We could in fact significantly *increase* and *decrease* the likelihood that a person would comply with an enforced social norm, fairness in this case, by manipulating (increasing or decreasing) the neural excitability of the LPFC. The neural networks embedded in this region will for sure deserve attention in future studies aimed at improving social behavior.



## **4. Pragmatic implications of empirically studying moral decision-making**

(Adapted from, Ugazio, G., Heinzelmann, N., and Tobler, P. N. (2012) *Frontiers in Neuroscience*.

### **4.1. Abstract**

When considering morality, at least three core questions come to mind: Which is the best normative theory? Which theory best describes moral decision-making? Why do people not behave the way they ought to behave? While the first question resides in the normative domain and concerns the way agents ought to make moral decisions, the second resides in the empirical domain and concerns the way we actually make those decisions. Both of these questions have been treated with some detail previously. Here we focus on the third question, which is a pragmatic one, reaching both into the normative and descriptive domains of morality. It naturally leads to another question, which we also cover: Can we narrow the gap between what people are morally required to do and what they actually do? We argue that two main problems usually keep us from acting and judging in a morally decent way: Firstly, we make mistakes in moral reasoning. Secondly, even when we know how to act and judge, we still fail to meet the requirements due to *akrasia*. We describe possibilities with which such shortcomings can be overcome as suggested by findings from neuroscience, economics, and psychology. Whether such possibilities should be implemented is a normative question that we put up for discussion.

### **4.2. Introduction**

A sharp distinction has been made between the descriptive domain of morality, i.e. the way agents behave or make moral judgments, and the normative domain, i.e. the way agents ought to behave or make moral judgments. In the empirical sciences, there has been an on-going debate about which theory describes moral decision-making best. Similarly, normative moral philosophy has been discussing which ethical theory is superior to the others.

However, whenever we watch the news or observe our social environment, both of these issues are of comparably little importance to us. The question that usually concerns us is not: How do people behave? Or: How ought they to behave? But rather: Why do they fail to behave in the way they should?

This last question is not purely an empirical one, as it involves an assumption about how one ought to behave. Nonetheless, it is neither an ultimately normative one, as it relies on empirically observable facts about human behavior. The issue is rather a pragmatist one, reaching both in the descriptive and the normative domains of morality. It naturally leads to another question: What can we do about the fact that people often do not behave in a way they are morally required to?

In this essay, we elaborate on the two related pragmatist issues and give an outline of how to resolve them. We argue that two main problems usually keep us from acting and judging in a morally decent way: Firstly, we make mistakes in moral reasoning. Secondly, even when we know how we ought to act and judge, we still fail to meet our obligations due to personal weaknesses.

#### **4.3. How ought we to act?**

Normative ethics tells us what we ought to do. The three most prominent contemporary theories are consequentialism, deontology, and virtue ethics (Tobler et al., 2008). There is no clear, simple and universally accepted definition for any of them; therefore we will give a brief account of how these concepts are understood in the present paper. Albeit rough and sketchy, we assume that these characterisations serve our present purpose well enough.

In one of its general forms, consequentialism tells us that the outcomes (consequences) of our actions ought to be as good as possible (Scheffler, 1988; Singer, 1993). There are numerous consequentialist theories which can be classified in various ways. Philosophers traditionally distinguish act and rule consequentialism. Act consequentialism holds that the outcome of single actions ought to be as good as possible. As consequences of single actions are often difficult to predict, attempts have been made to facilitate the decision process of an agent. In this vein, rule consequentialism focuses on action-guiding rules, claiming that the consequences of the rules be as good as possible. Actions are then evaluated with respect to these rules.

Also, different consequentialist approaches disagree on what the goodness of an outcome consists of. The most popular one, utilitarianism, holds that we ought to do what increases people's happiness or, decreases their unhappiness. Hereby, the good of everyone has to be taken

into account and everyone's good counts equally. We ought to act in a way that maximises the good of all and in no other way. Jeremy Bentham (1996/1789), one of the founders of classical utilitarianism, argued for a felicific calculus that allows measuring the outcome of various actions, i. e., the pleasure these actions may produce. Such a method presupposes that all pleasures are comparable and quantifiable and that they are, as consequences of an action, to greater or lesser certainty predictable. After such hedonic approaches to (experienced) utility had been largely abandoned by economics, they have more recently been taken up again by behavioral economics (Kahneman et al., 1997). Moreover, some formal treatments of welfare economics (Harsanyi, 1955) and prosocial preferences (e.g. Fehr & Schmidt, 1999) also have consequentialist roots.

“Deontology” is a collective term denoting a variety of theories which, from a linguistic point of view, assign a special role to duties, as “deontology” refers to the study or science of duty (deon = duty). Deontology requires us to fulfil our moral duties but such a general claim is also made by consequentialist theories, which hold that it is our moral duty to act in such a way that the outcomes be as good as possible. Therefore, deontology is sometimes identified with non-consequentialism, the claim that the wrongness or rightness of an action is not only determined by the badness or goodness of its consequences. For instance, an action can be assigned intrinsic value because of the agent's willingness that the principle - or maxim - on which the action is performed should become a universal law, a criterion established by Immanuel Kant (1965 /1785). Kant's ethics and the theories derived from them are often seen as prominent candidates of deontology. Another central requirement of Kant's ethics is to never treat a human being as a means to an end. Thus according to Kant and in contrast to consequentialism, it would be morally wrong to let one person die if thereby two other human lives could be saved.

In this text, we will schematically conceive of deontology as a rival both to consequentialism and to virtue ethics. Virtue ethics usually goes beyond the question of what we morally ought to do. This has historical reasons: The earliest prominent account of virtue ethics has been developed by Aristotle (Roger, 2000) who was concerned with the best way for a human being to live, where “best” is not to be understood as “morally best”. A central claim of contemporary virtue ethicists is that living virtuously is required in order to flourish. Roughly

speaking, a virtue is a disposition to act appropriately for the right reason and thus requires practical wisdom. Flourishing can be described as living fulfilled and happily, which goes beyond mere momentary subjective well-being but refers to an overall outlook and life as a whole. All of these theories are primarily concerned with the question of how we ought to act but rarely consider how individuals actually do behave. We shall turn to this topic in the following section.

#### **4.4. How do we act?**

Empirical research on human moral behavior has focused primarily on two topics: action and judgment. As these two aspects of moral behavior have been studied using rather different approaches, we will treat each of them separately here. First, we consider the literature studying the effects of norms on people's actions (Bicchieri, 2006; Gibson et al., 2011). Second, we shall focus on the literature studying the psychological mechanisms underlying moral judgments (Mikhail 2007; Hauser 2006; Prinz 2006; Moll et al. 2005; Greene et al., 2001).

From a wider perspective, the question arises whether moral judgment translates into moral behavior. This issue is controversial and has received a variety of answers (e.g. Schlaefli et al., 1985). One view (Bebeau et al., 1999) suggests that a moral act requires not only that an agent judges one course of action as moral but also that one identifies a situation as moral (e.g. that consequences of distinct courses of action have differential welfare implications; defined in Bebeau et al. 1999 moral sensitivity), chooses the moral over other courses of action (defined in Bebeau et al. 1999 moral motivation) and persists to implement the goal of the action (defined in Bebeau et al. 1999 moral character). In this view, it would be expected that judgment and action are positively but weakly correlated, which seems to be the case (Blasi, 1980).

##### **4.4.1. Moral Action**

One of the most successful approaches to study moral action has been to observe how people's behavior changes depending on the saliency of a norm. Scholars working in this field developed several models to show how the utility assigned by a person to different outcomes in a given situation is modified by the presence of a norm. Norms motivate compliant behavior mainly in two ways: (a) they modify the expectations an individual has regarding others'

behavior (Bicchieri, 2006), and (b) they generate a personal cost for violating the action course prescribed by the norm (Gibson et al., 2011).

While Bicchieri's work focused mainly on providing a theoretical description of how and when social norms are most likely to emerge and influence individual behavior, other scholars provided empirical evidence demonstrating the influence of norms on behaviors in a social context. For instance, recently Gibson and colleagues (2011) tested the influence of the moral obligation of being honest (or not lying) on individuals' behavior in an economic context. The authors tested the hypothesis that when being incentivized to lie by being able to make a greater profit through not telling the truth, the willingness of an individual to behave immorally, i.e. to lie, was correlated with the importance one assigned to being honest. More specifically, those individuals attributing high importance to the honesty norm were extremely insensitive to the cost of telling the truth, which suggests that the moral value of respecting a moral obligation (of being honest) can outweigh economic costs of respecting it and even prevent cost-benefit trade-offs altogether. Her works, to now, have solely proposed theoretical model of how norms may diffuse in a society but have not been yet experimentally tested.

#### **4.4.2. Moral Judgment**

In this brief section the moral descriptive theories proposed in moral psychology, thoroughly introduced both in the general introduction and in the introduction to the first study, are briefly summarized in this context to help the reader visualize the existing parallels between these descriptive theories and the normative theories proposed in the philosophical literature.

Whereas psychological research on moral judgments has captured them predominantly as a cognitive, controlled process and focused on moral development in the 20th century (Kohlberg 1976; Piaget 1932), it has in recent years mainly developed around two research questions: a) do moral judgments stem from intuitions or from conscious reasoning, and b) which psychological processes are involved in moral intuitions (Cushman et al., 2010). Roughly, we can distinguish four different approaches to these questions.

From a first perspective, following Hume's idea that moral judgments result from "gut feelings" (Hume, 1777/1960), some scholars proposed that moral judgments predominantly result from intuitions of an emotional nature (Prinz 2007 and 2006; Woodward & Allman, 2007).

Second, others agree that moral judgments indeed stem from intuitions but they deny that such intuitions are of emotional nature, arguing instead that moral intuitions are the product of moral specific psychological mechanisms named "universal moral grammar" (Huebner et al., 2008; Mikhail, 2007; Hauser, 2006). According to this view, neither conscious reasoning nor emotions play a causal role in determining moral judgments, suggesting that these two processes actually occur after the moral judgment has been produced by the "moral grammar" mechanism. Such position seems to be in line with Rawls' hypothesis that humans possess a "moral instinct" (Rawls 1971).

From a third point of view other scholars put forward a dual-process theory of moral judgment (Greene et al., 2004) suggesting that moral judgments result from two psychological mechanisms: emotions and conscious reasoning. It is consequently claimed that different moral judgments are underpinned by different psychological systems (Cushman et al., 2010).

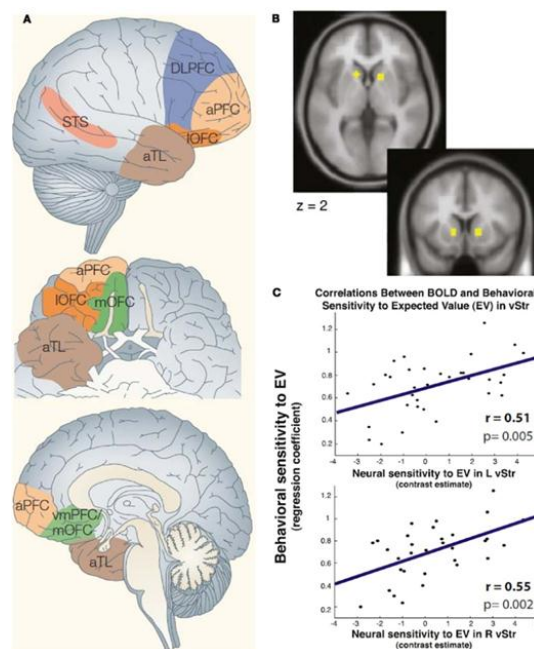
Finally on a very similar stance, a fourth theory acknowledges that moral judgments rely on multiple psychological mechanisms, and therefore that both emotions and conscious reasoning play a role in moral judgments. However, in contrast with the third view described above, it is argued that different moral judgments are not underpinned by different psychological systems, but rather that all moral judgments will involve cognitive and emotional mechanisms in competition against each other when a moral judgment is produced (Moll et al., 2005 and 2008).

#### **4.4.3. Neural Underpinnings**

The advent of neuroimaging methods allowed studying the intact brain of healthy volunteers while they make moral judgments and decisions. This line of research has identified a variety of brain regions that are active during moral cognition (see Fig. 11 and, for review, see: Forbes & Grafman, 2010; Moll et al., 2008 and 2005; Raine & Yang, 2006). These regions include the prefrontal cortex, particularly ventral, medial, dorsolateral and frontopolar sub-regions, posterior cingulate cortex, anterior temporal lobe, superior temporal cortex, temporo-

parietal junction, striatum, insula, and amygdala. Many of these regions are also implicated in “theory of mind” tasks requiring consideration and inference of others’ thoughts and desires (Bzdok et al., 2012) and impaired in patients with antisocial disorders, in agreement with the notion of impaired moral decision making (see Fig. 12; Raine & Yang, 2006).

One could next ask whether neuroimaging can contribute to informing theories of moral decision-making. Could it help deciding between the different theories outlined in section 5.4.2 (even though some of them may not be mutually exclusive)? Or, more specifically, can neuroimaging inform us about the degree to which emotions are involved in moral judgment?



**Figure 11.** Brain regions implicated in moral judgment and decision making (Figure 11a originally published by Moll et al. 2005 as figure 1a; Figure 11b and c originally published by Shenav & Greene 2010 as figures 5c and 5d). (a) Cortical regions. Note that the posterior cingulate cortex and the angular gyrus (temporoparietal junction) have also been implicated in moral judgments (shown in Figure 12). aPFC: anterior prefrontal cortex; aTL anterior temporal lobe; DLPFC: dorsolateral prefrontal cortex; IOFC: lateral orbitofrontal cortex; STS: superior temporal sulcus; vmPFC: ventromedial prefrontal cortex. Adapted with permission from Moll et al. (2005). (b, c) Example for striatal involvement in moral decision making. The task employed moral dilemmas. In each trial, subjects rated how morally acceptable it was to save a group of individuals from death with a known probability rather than a single individual with certainty. Across trials, group size and probability varied. Group size and probability should be multiplied to compute the expected number of lives saved. (b) Regions in ventral striatum previously identified by Knutson et al. (2005) as processing reward value. (c) In the regions shown in (b), individual neural sensitivity (contrast estimates in activation increases) correlated with behavioral sensitivity (beta estimates in rating) to the expected number of lives saved.

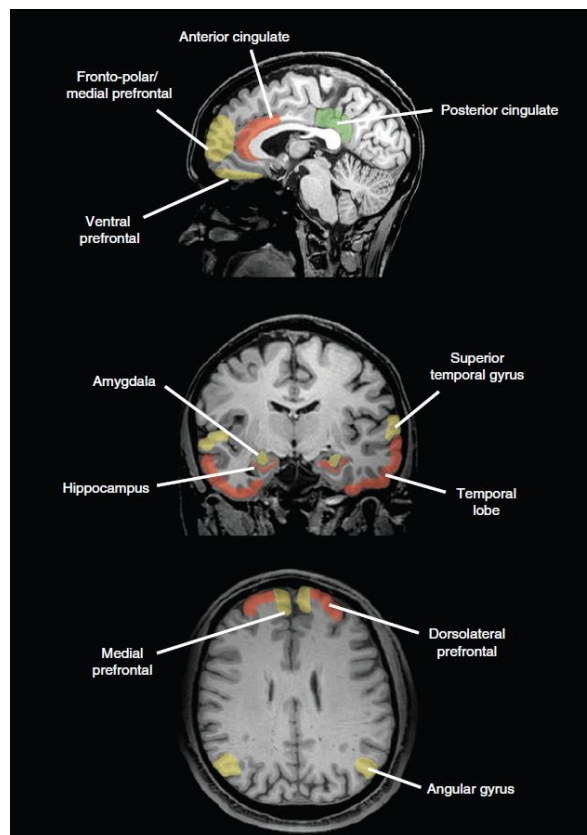
When asking such questions one is often tempted to make reverse inferences from brain activation to mental function. However, given that most brain regions contribute to more than one function, such inferences are at best probabilistic (Poldrack, 2006; Poldrack, 2011). Moreover, they are limited by the response specificity of the brain region under study and by the precision with which mental functions are parsed conceptually and assessed empirically (Poldrack, 2006). Nevertheless, some attempts to answer those questions have been made.

For example, an extension to Hume's view mentioned above may be suggested by the involvement of dorsal and lateral frontal regions in moral judgment (e.g. Greene et al., 2001). This would be based on the notion that these regions play a stronger role in more deliberate, goal-directed and cognitive than automatic and emotional functions (Forbes & Grafman, 2010). Moreover, all of the regions implicated in moral judgment have been implicated also in other mental functions. This seeming lack of evidence for a neural substrate exclusively devoted to moral functions (Young & Dungan, 2012) does not support the universal moral grammar approach; if one assumes that moral functions have evolved from non-moral functions or that the mental functions required for other types of judgments can be used also in the moral domain (Tobler et al., 2008) it is perhaps not surprising that so far no region has been singled out as a uniquely moral center of the brain. In principle though it is still conceivable that such a substrate may be revealed by finer grained methods, such as single cell recordings. However, it is more likely that what may be identified in future studies is that there may exist a moral-specific neural network, relying on several brain structures each of these contributing to the moral decision. .

Neuroimaging and lesion work also point towards a role for emotion in moral judgment. The ventromedial prefrontal cortex (VMPFC) is involved in emotion processing and also activated when a subject makes moral judgments (reviewed in Young & Koenigs, 2007). Lesions of this region result in blunted affect (hypo-emotionality) as well as increased emotional reactivity to environmental events (Anderson et al., 2006). Activations are increased by pictures with moral emotive content (depicting e.g. abandoned children, physical assaults) compared to pictures with non-moral emotive content of similar emotional valence and sociality (Harenski & Hamann, 2006; Moll et al., 2002) and by moral compared to semantic judgments (Heekeren et al. 2003, 2005). Patients with lesions of the VMPFC are more likely than controls to endorse harming someone in order to benefit a greater number of other people (Thomas et al., 2011;



Ciaramelli et al., 2007; Koenigs et al., 2007). In healthy subjects the strength of skin conductance responses to such moral dilemmas correlates inversely with the propensity to endorse harm for the greater good (Moretto et al., 2010). By contrast, VMPFC patients fail to generate such emotive responses before endorsing harm (Moretto et al., 2010). Thus, at least some moral judgments appear to be caused by emotions.



**Figure 12.** Comparison of brain regions preferentially activated during moral judgment and decision making (green), regions impaired in patients with antisocial disorders such as antisocial personality disorder and psychopathy (red) and common regions (yellow). One possible interpretation is that emotions as underpinned by the common regions prevent breaking of moral rules, the defining deficit of antisocial personality disorders. The angular gyrus lies at the junction of temporal and parietal cortex. Raine and Yang (2006, figure 1).

Although much of the literature has focused on prefrontal cortical regions, moral judgment and decision making are clearly not a purely prefrontal or, more generally, neocortical matter. Activation in the striatum, for example, is affected by the moral status of a partner with whom one performs economic exchanges (Delgado et al., 2005) and reflects behavioral sensitivity to the “moral expected value” (number of lives saved) of moral actions (Shenhav & Greene, 2010; see Fig. 11 b). Based on its general role in action selection (Balleine et al., 2009),

one would also expect the dorsal striatum to contribute to the selection of moral actions. The amygdala contributes to the learning of fear and distress experienced by others (Blair, 2007; Olsson et al., 2007); empathy-induced insula activation correlates with subsequent prosocial behavior (Masten et al., 2011). Thus, although these regions may primarily serve different functions they can nevertheless be harnessed for moral judgments and decisions.

#### **4.5. People do not behave in a way they ought to**

Combining insights from the two previous sections, this part of the paper will establish the claim that human beings often do not behave in a way they ought to. Although it is clear that discrepancies can arise from a variety of issues, including moral sensitivity, judgment, motivation and character, we will concentrate on two more recently discussed phenomena: cognitive biases and emotional influences. Both these phenomena are morally problematic in that they reflect the influence of morally irrelevant features on actions and judgments. We will briefly clarify this point for each of the three ethical theories outlined in section 5.1. As mentioned before, consequentialism requires that only the ultimate consequences of an action or judgment are relevant to its moral evaluation. Therefore, features such as the emotional state of the agent or the framing of several options to choose from are not to be taken into account. However, as we shall elaborate in the following, there are a variety of instances in which agents are influenced by such cues and therefore do not act and judge in a morally decent way.

From a deontological point of view, a morally right action or judgment is to be made from duty, that is, out of reverence for the moral law. Accordingly, any other feature of a situation, such as the agent's uneasy feeling towards the morally prescribed action course, is to be ignored. However, empirical evidence will be given below that individuals often fail to meet this normative requirement. Virtue ethics outlines the character traits which distinguish a virtuous person. Amongst them are the faculty of practical reasoning and specific virtues such as justice or temperance. There is, however, solid evidence that agents frequently fail to display these traits in their behavior and judgments, as this section will make clear. In the following, we will show in greater detail in what ways individuals are biased or influenced by their emotions. For some cases, we will exemplarily explain how the actions and judgments in question are morally dubious from a deontological, consequentialist, or virtue ethical perspective.

#### **4.5.1. Biased behaviour**

Briefly, a cognitive bias is an unconscious tendency to judge a certain element in a way that depends on one's own preferences, expectations, and experiences. Cognitive biases are similar to perceptual biases such as optical illusions (e.g. the Müller-Lyer illusion, 1889). Instead of influencing our perceptual skills, cognitive biases affect people's cognitive capacities. We will give some examples for this phenomenon below.

Firstly, a known cognitive bias that strongly affects moral actions is the so-called bystander effect, i.e. “the more bystanders to an emergency, the less likely, or the more slowly, any one bystander will intervene to provide aid” (Latane & Darley, 1968 p.1). Latane and Darley (1968) recreated an emergency situation in the lab in order to test the reactions of participants. The higher the number of bystanders, the lower the percentage of participants who decided to intervene and the longer the time it took them to do so. The presence of others affects one judgment by easing the pressure to help generated by the person in need by spreading the sense of responsibility over the participants: if I am the only one there to help a person, the pressure to rush and help him falls only over me, but as the number of the people present increases the weaker this pressure is felt by each of the individual. Moreover, another cognitive bias affecting one's judgment in this situation is the need for conformity: given that this situation is not a typical one, most of the bystanders would not be confident on what the appropriate action course is. Therefore most will wait to see what the others do (Manning et al., 2007). However, such behavior is morally questionable. For instance, from a deontological perspective, it is highly plausible to assume that an agent has a strong duty to help a victim in an emergency. Besides, such a duty is often legally prescribed, i.e., non-assistance of a person in danger is widely regarded as tort. The presence of bystanders and their number does not relieve the agent from his moral duty. Failure to act from the duty to help is thus a severe moral transgression from a deontological point of view.

Secondly, the next cognitive bias taken into consideration here is the one known as the identifiable victim effect (Schelling, 1968; Redelmeier & Tversky, 1990): one is more likely to help a victim if he is easily identifiable. An example of this behavior is people's widespread inclination to save one little child from drowning in a shallow pond but to refrain from making a

small donation that would save twenty-five children from starving to death in Africa (Hauser, 2006). This pattern of results was consistently found in numerous previous studies observing people's behavior in similar situations (Viscusi, 1992; Whipple, 1992; Redelmeier & Tversky 1990; Calabresi & Bobbitt 1978). This behavior seems to be driven by several cognitive elements, for instance: people tend to value the situations based on proportions, therefore the drowning kid is more likely to be helped because 100% of the population in this situation (given he is alone) will die, while the kid in Africa starving is one among the whole African population; furthermore, one tends to base his decision given the immediateness of the emergency: the kid drowning kid can be saved instantly, while to save he kid in Africa will take days before one can actually deliver his help. Again, this is morally dubious behavior, as we will argue from a virtue ethicist's viewpoint. Generally, charity and justice (or fairness) are regarded as moral virtues. Assume further, plausibly enough, that the overwhelmingly important point about being charitable is the benefit of the person receiving aid. Then a virtuous agent seems to be morally obliged to helping both the drowning child and the starving kids. Helping one but not the others would amount to a failure of exhibiting charity and justice and is therefore a morally reprehensible action. From a consequentialist perspective it could be argued that saving twenty-five is likely to have better consequences than saving one. Thus, failing to save the larger number would presumably be morally dubious also from a consequentialist perspective.

#### **4.5.2. Emotionally influenced behaviour**

Among the elements influencing moral behavior, emotions play an important role. Although, as for cognitive biases, people are usually unaware of the influence that emotions have on their behavior, several studies have shown that brain areas associated with emotion are involved in various decision making tasks, including the formation of moral judgments. A seminal study by Greene and colleagues (2001) has shown that emotions are usually sensitive to the means used for an action, while cognitive processes are sensitive to the consequences resulting from this action.

Other studies investigating the role of emotions in moral judgment showed that moral condemnation of an event (i.e. how wrong you think something is) is strongly influenced by the emotional state of the person evaluating it. Haidt and colleagues (Eskine et al. 2011; Schnall et

al., 2008; Wheatley & Haidt, 2005) ran a series of studies which showed that induced disgust can yield harsher condemnations of a set of disgust-related moral violations such as incest. Recently, we (see section 3 or Ugazio et al., 2012) have provided evidence that when a person judges a moral scenario, different emotional states will influence her choices in opposite ways. People who were induced to feel anger were more likely to judge a moral action in a permissive way compared to people in a neutral emotional state, and people induced to feel disgust were more likely to judge the same actions in a less permissive way.

The influence of emotional states on moral judgments and actions, in particular if the emotions stem from morally irrelevant factors of the situation, is morally problematic from all the moral theories outlined in section 1. Consider consequentialism first and recall that from this perspective, the only aspects relevant to a moral evaluation are the outcomes of an action, decision, etc. In particular, the emotions of the agent are only relevant to the extent to which they are part of the overall utility affected by the outcomes. Hence a judgment or action crucially influenced by the feelings of the agent is morally wrong from a consequentialist outlook. According to deontology, a morally right action or judgment is to be performed out of duty. Kant (1785/1959) famously declined that an action out of inclination fulfills this criterion. As emotions are regarded as inclinations of this sort, a judgment or action highly influenced by an emotion cannot be morally right. From the point of view of virtue ethics, the actions and judgments described in this section seem to be morally questionable because they violate the virtue of temperance. A virtuous person is supposed not to be dominated by his passions. However, the extent to which temperance is violated will presumably depend on the degree to which the action or judgment is influenced by the emotions.

Having given evidence for the claim that individuals often do not behave and judge in a morally sound way, we shall in the following section provide details on what we believe are the most important reasons for these failures.

#### **4.6. Why do we not behave in morally decent ways?**

A first step towards a solution to the problem that people often do not behave in morally decent ways consists in analyzing the reasons and mechanisms of this behavior. Our hypothesis

is that we do not behave in a way we ought to either because we do not know what to do or because we fail to carry out the right action despite our better knowledge.

For the first problem – we make mistakes in moral reasoning – a range of different causes can be given. The most obvious one is a lack of cognitive capacities. For instance, we suddenly find ourselves to be free-riders on a train because we simply forgot to validate our ticket. In this case, it is simply bad memory, lack of planning, distraction or time pressure that led us to a moral transgression. Inappropriate moral decision-making may also occur as a consequence of people's ignorance of important information. Such ignorance then prevents them from drawing the correct conclusion how to act. For example, a consumer who wants to support fair working conditions may make a wrong decision because he is not aware that the company selling the product he chooses has recently been found guilty of exploiting sweatshop labor. In addition, defective moral reasoning may be behind such phenomena as cognitive biases. Consider, for instance, the identifiable victim effect as described in the previous section. It seems to stem from a lack of reflection on the two scenarios, their comparison and moral evaluation, ultimately leading to the violation of the virtues of justice and fairness.

In other words, the first reason we identify as compromising our capacity of behaving in morally decent way is that we are not able to control our moral thinking in such a way to avoid being influenced by morally irrelevant properties of circumstances: being surrounded by others leads us to refrain from helping someone in need we would normally help in the case where less, or no others were witnessing the situation; not being able to identify a victim interferes with our valuation inducing us to underestimate the real needs of this person, while on contrast being able to identify the victim could lead us to overestimate such needs helping more than needed. The second problem – despite knowing how we ought to act, we fail to carry out the right action – can be analyzed in a variety of ways. We will consider only a selection here. Failure to act in a way that has been acknowledged of being the morally correct one may be due to personal weaknesses. The most prominent one is *akrasia*, sometimes also described as weakness of the will (Kalis et al., 2008). A person is called *akratic* if one acts against his/her own standards or aims. Succumbing to some temptation, e.g., eating another portion of ice-cream despite your knowing you are thereby taking away someone else's share is usually regarded as an *akratic* action (Austin, 1961). The concept of *akrasia* depends heavily on the underlying idea of man. If

we share Socrates' view of a completely rational homo-economicus, akrasia simply does not exist. Similarly, Aristotle and Aquinas have regarded akrasia as a result of defective practical reasoning whose result is a morally bad action (see also Hare, 1981 and 1963; Davidson, 1970). However, if we believe that akrasia goes beyond fallacious reasoning, the difficult question arises of what akrasia actually is. Some have claimed that it is a conflict of competing forces, for instance, according to Augustine, between incompatible volitions. Others have described it as an instance of self-deception (Wolf 1999; Schälike 2004). In an Aristotelian vein, Beier (2010, see also 2008) argues that it is a result of underdeveloped virtues, that is, a defect in character building.

As far as we know, the philosophical concepts and theories concerning akrasia and related phenomena have not yet been linked to empirical research on defects of self-control, empathy and self-involvement. Such an enterprise might, however, provide fruitful insights for both approaches. As the literature on behavioral and neuroscientific research is vast, we will confine ourselves to a very brief review of evidence concerning self-control here. An action out of self-control is generally defined as the choice of larger-later rewards over smaller-sooner ones (Siegel and Rachlin, 1995). Self-control has also been defined as the regulation of habits. From another perspective, self-control amounts to the control of emotional reactions (Ochsner and Gross, 2005). Both the second and the third approach regard self-control as a control of automatic reactions involving similar neural circuits. Neuroscientific research investigating the brain areas involved suggests that the dorsolateral prefrontal cortex (DLPFC) modulates the value signal encoded in the ventromedial prefrontal cortex (vmPFC) which in turn drives choices and decisions (Hare et al., 2009). The DLPFC promotes task-relevant processing and eliminates irrelevant activities. Future research into DLPFC and its interactions with vmPFC and other brain regions may shed new light on how to analyze self-control and akrasia and how to influence those phenomena.

In sum, the philosophical conception of akrasia may be linked to a lack of self-control in the following way: relying on Beier, akrasia can be regarded as defective character building which essentially involves the development of self-control. This, in turn, will yield agents falling prey to morally irrelevant aspects of a situation, such as cognitive biases or emotional influences, which affect behavior and judgment. To illustrate, consider an example from the previous

section: depending on their emotional states, subjects regarded moral transgressions more or less severe (see section 3.1, or Ugazio et al., 2012). That is, they could not separate their feelings from a consideration of a moral scenario which amounts to a defect of control over the emotions.

Another issue that hinders us from acting in morally decent ways may be certain character traits. A general reason for both morally fallacious reasoning and failure to carry out the action identified as the right one is the evolutionary background of human beings. Morality can be viewed as a product of the phylogenetic history of our species which has evolved in an environment different from the one we live in today. More precisely, it is commonly believed that reciprocity became a part of moral behavior because it enhanced the evolutionary fitness of reciprocating individuals (reciprocal altruism, Trivers, 1971). Similarly, pro-social behavior within a group increased the reproductive abilities of its members in comparison to no- or anti-socially behaving groups (group selection, Sober and Wilson, 1998). Likewise, altruistic behavior towards one's own kin may increase the likelihood of spreading the shared genes (kin selection, Hamilton, 1964).

To give some examples for evolutionary explanations of moral behavior, immediate and strong emotional reactions to a given situation probably evolved because they facilitate a quick reaction which in turn improved survival, for instance the fight-or-flight response to predators. The theory of kin selection can explain why humans evoke emotional reactions such as caring love towards their offspring and may favor them over foreigners: by helping the former and not the latter, their own genes are more likely to be passed on in the future. Likewise, we are now equipped with biases that automatically and unconsciously guide us in a way that helps to spread our genes. For instance, the identifiable victim effect increased the safety of the young in the agent's close environment who shared genes with him with higher probability than did children further away. According to group selection, such biased behavior also improved the evolutionary fitness of one's own group, as helping close-by group members rather than faraway out-group individuals would favor one's own group and eventually the agent himself.

Related to this point, reasons for why we do not behave in morally decent ways can be regarded from a cultural perspective. On this view, morality can be seen as a relatively recent development, crystallized in laws and rules for social conduct. In this vein, the philosopher



Friedrich Nietzsche has argued against moral systems such as Kantian, Christian, and Utilitarian ethics, criticizing that these codes of conduct are “detrimental to the higher men” (Nietzsche, 1966, p. 228) while benefiting the “lowest”. From a similar perspective, morality may be seen as a fear of punishment which evolved originally and is exploited by legal systems. In this view, failures of morality arise whenever people do not experience enough fear of punishment. Presumably, the lack may come from the person or the situation. Empirical research proves helpful to investigate and explain each of the problems mentioned, providing a basis on which we can search for solutions. We will turn to this topic in the following section.

#### **4.7. Improving moral behaviour**

Having provided evidence (see section 5.5) that people often make inconsistent, if not mistaken, moral decisions and possible explanations for such irrational behaviour (see section 5.6), in this section we discuss possible means by which improving humans’ moral decision capacities, particularly via nudging, training, pharmacology, and lastly brain stimulation.

##### **4.7.1. Nudging**

A nudge has been defined as an “aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options” (Thaler and Sunstein, 2008, p. 6). Other than regulating, nudging does not eliminate possible courses of action. For example, a school canteen can increase pupils' intake of vitamins by placing fruit salad or similar desserts in front of the chocolate cakes and sweets. This would be a nudge, whereas banning all alternatives to a healthy dessert would be a regulation. Nudging makes use of inclinations and biases, e.g., the fact that people tend to favor items displayed at eye level or often eat the portion they are served regardless of its size. Marketing strategies have benefited of these insights long ago, relying on long lasting research projects into consumer habits and psychology.

Nudging has mainly been investigated as a means to tackle population health issues, such as obesity and addiction to alcohol, nicotine or other substances (Downs et al., 2009; Just and Payne, 2009; Zimmerman, 2009). However, it can be equally relied on in order to approach moral issues: it provides paternalistic institutions with strategies to succeed in guiding their clients, patients or charges to the morally right decisions or actions (Thaler and Sunstein, 2003).

For instance, given the assumption that organ donation is a morally praiseworthy action, a government can yield an increase in organ donors by making the donation of organs the default option of which you have to opt-out if you do not want to be a potential organ donor.

However, nudging in moral contexts raises a lot of issues. First, it is questionable whether a morally praiseworthy action loses its praiseworthiness if it had not been performed without the relevant nudge. This depends on whether an action is to be evaluated only on the basis of its results or also with regard to the states of mind of the agent. Second, as nudging itself is neither morally good nor bad, the question arises how it could help us to improve moral decision-making and acting at all. Nudging may well be abused by the nudger for his personal interests. Third, the practice of nudging itself may be questioned on the ground of fear for autonomy and respect. These and other questions will be discussed in section 6. For now, we shall outline some further means and methods that might be useful for an improvement of moral practice.

#### **4.7.2. Training**

Although already Aristotle suggested that sound judgment needs practice (Roger, 2000), there is little empirical research on direct training of moral decision-making. In as far as it is feasible to train cognitive and emotional functions and such training transfers to other domains it may also be conceivable to improve moral decision-making indirectly by training these functions. Working memory performance increases with training techniques such as an adaptive dual n-back task (e.g. Jaeggi et al., 2008), or an adaptive order-and-location memory task (e.g. Klingberg et al., 2005; Thorell et al., 2009). Working memory training transfers to other domains, including fluid intelligence (e.g. Jaeggi et al., 2008), attention (Thorell et al., 2009) and response inhibition, at least in children with ADHD (Klingberg et al. 2005). However, transfer appears to occur primarily in closely-related domains (Li et al., 2008) and only in individuals in which initial training is successful (Jaeggi et al., 2011).

Response inhibition can be trained with go/no-go and flanker tasks (Thorell et al., 2009) whereas executive attention improves after training with a battery of anticipation and stimulus discrimination exercises (Rueda et al., 2005) but training effects seem to transfer less readily than with working memory training. Based on the hypothesis that utilitarian components of moral decision-making depend more on cognitive factors than deontological ones (Greene et al.,

2001), one may speculate that training cognitive factors would improve specifically utilitarian components of moral decision-making. However, given that transfer appears to be limited to closely-related domains, it is questionable whether moral behavior would benefit from such training.

Training of emotional factors can improve aspects of moral decision-making. For example, a Buddhist compassion-enhancing technique increases provision of help to another player in a virtual treasure hunt game (Leiberg et al., 2011). In the same game, the duration of compassion training correlates with helping particularly in situations in which the other player cannot reciprocate help. By contrast, compassion training does not affect giving money to others in a dictator game, where subjects decide how to split an amount of money assigned to them between a stranger and themselves (Leiberg et al., 2011). Taking these findings further, one may wish to investigate whether deontological components of moral decision-making are influenced more by emotion training than utilitarian components. Through increasing effort-levels required for achieving reinforcement as well as exercises such as monitoring and improving posture, trying to improve mood states, and monitoring eating, self-control can be increased in humans and rats respectively (reviewed in Strayhorn, 2002). Accordingly, it has been proposed that self-control acts like a muscle that can be trained or fatigued depending on experience (Baumeister et al., 1994). Insofar as self-control reflects a virtue, self-control training may be beneficial from a virtue ethics perspective.

#### **4.7.3. Education**

Moral education has a long tradition and received consideration from all three philosophical theories introduced above (Althof & Berkowitz, 2006). It largely follows on from the (deontologically flavored) views of Piaget (1932) and Kohlberg (1984) and focuses primarily on the development of moral reasoning. By contrast, the related character education has a stronger grounding in virtue ethics and utilitarianism and aims to promote moral actions leading to good consequences in educated citizens (Althof & Berkowitz, 2006).

Within a “Kohlbergian” framework, interventions specifically designed to promote moral education are more effective than control interventions or the passage of time (Schlaefli et al., 1985; King & Mayhew, 2002). Moreover, longer term (up to 12 weeks is optimal) interventions

that focus on peer discussion of moral dilemmas, thereby leading to practice in moral problem solving, and interventions that focus on personality development and self-reflection are more effective than shorter-term interventions ( $\leq 3$  weeks) and interventions that focus on academic content such as criminal justice, law and social studies (effect sizes: 0.36-0.41 versus 0.09; Schlaefli et al., 1985). Treatment effects are more pronounced in older ( $\geq 24$  years old) compared to younger (13-23 years old) subjects, although this may be partly due to selection bias (older subjects are more likely to be volunteers) or other methodological issues. Although the effect sizes of interventions are small to moderate, they lead to 4-5 years of natural growth compared to no intervention (Schlaefli et al., 1985), suggesting that education may be a promising avenue for future research.

#### **4.7.4. Pharmacological enhancement**

The field of cognitive enhancement by pharmacological means has received attention in recent years (reviewed e.g. in Jones et al., 2005; Illes & Sahakian, 2011) but the first empirical investigations have focused primarily on improving cognition as such, rather than on moral decision-making. Below, we review a few example studies with a more direct link to moral behavior. Before going further though, it is important to note a few caveats: (1) It is not necessarily the case that more of a given pharmaceutical agent results in monotonic increases in function. Instead, at least some functions may require an intermediate level of the agent. Increases beyond that level result in decreases in the function. An example for this notion comes from working memory and dopamine (reviewed e.g. in Cools & D'Esposito, 2011). (2) Individual differences can moderate the relation of how pharmaceutical agents affect function. Such individual differences can be genetic or psychosocial. An example comes from the Taq1A DRD2 (dopamine D2 receptor) gene, where the presence of an allele (A1+) is associated with reduced dopamine receptor concentration, decreased neural responses to reward, but enhanced neural reward responses after delivery of a D2 receptor agonist compared to A1- subjects (Cohen et al. 2007). The endeavor of improving a given function may thus require tailoring agents and dosage to individuals. (3) Improvements for one function may come at the expense of costs for another. For instance, improvements in social functions may come at a cost of reduced cognitive functions. In other words, considering that the brain has limited computational resources, if more are allocated to social aspects, this has to have a cost which might result in a diminished

cognitive capacity, say for instance in counting (see for instance Hebb's rule on neural plasticity, Hebb, 1949). Ethical questions become pertinent in this case in that one would have to argue why one function is ethically more important than another. (4) Pharmaceutical agents administered systemically act in a sustained fashion over time but the relevant functions may be implemented in a temporally more phasic fashion. Moreover, the same pharmaceutical agent may have different functions at different time-scales (e.g. Fiorillo et al., 2003; review in Schultz, 2007).

Intranasal administration of oxytocin (24 international units) increases trust in the trust game (Kosfeld et al., 2005). More specifically, the average initial amount passed by an investor to a trustee is 17% higher under oxytocin (45% of participants showing maximal trust) than under placebo (21%). Proposers' offers are also enhanced by oxytocin in the ultimatum game (Zak et al., 2007). By contrast, non-social risk taking, trustworthiness of trustees (the amount returned by trustees) and amounts offered in the dictator game remain unaffected by oxytocin, excluding less specific effects on risk perception and pro-sociality more generally. Thus, oxytocin enhances an emotional aspect of moral behavior.

The administration of a selective serotonin reuptake inhibitor (30 mg Citalopram) increases the propensity with which people judge harming others as forbidden, if the inflicted harm is personal and emotionally salient (Crockett et al., 2010a). Moreover, it reduces the rejection of unfair offers in the ultimatum game (Crockett et al., 2010a; the rejection of unfair offers harms the proposer). Thus, serotonin may facilitate prosocial behavior or moral judgments more generally by enhancing aversion to harming others.

In sum, although one may claim that the side-effects of increasing trust or reducing unfairness punishment might lead to undesired effects, pharmacological manipulations may be used to increase certain traits that seem to be beneficial for moral behavior.

#### **4.7.5. tDCS/TMS**

As transcranial direct current stimulation has already been described in detail in a previous section (see Study II, section 3.3), we will briefly introduce TMS before providing examples of how these two stimulation methods have been used to influence behavior.

TMS (transcranial magnetic stimulation) is a technique of non-invasive brain stimulation which uses magnetic impulses to generate weak currents in specific brain regions. So far, two types of TMS have been used, single pulse TMS and repetitive TMS (rTMS). The first type of stimulation affects neural excitability similarly to anodal tDCS, resulting in a depolarization of the neurons targeted by the magnetic impulses. Such depolarization then results in the generation of action potentials in the stimulated neurons. By contrast, rTMS lasts much longer than single pulse stimulation. Therefore rTMS can increase or decrease the resting membrane potential of the stimulated brain region, depending on the intensity and frequency of the stimulation and on the coil orientation (Fitzgerald et al., 2006).

Using both these techniques scholars have shown that it is possible to directly manipulate social and non-social behavior in several tasks including temporal discounting (Figner et al., 2010) and norm compliance (see section 4 or Ruff et al., submitted). The latter study focused directly on moral behavior (i.e., complying with behavior prescribed by a norm). Other studies investigated processes which are related to moral behavior such as contributing to the enforcement of a fairness norm by costly punishing defectors, or mechanisms involved in shaping individuals' impulsivity. More specifically, Knoch and colleagues (2008) tested the role of dorsolateral prefrontal cortex (DLPFC) in punishing unfair behaviors. Measuring the altruistic punishments (Fehr & Gächter, 2002) responders inflicted to unfair proposers while playing an ultimatum game (Andreoni et al., 2003), the authors showed that reducing excitability by means of cathodal tDCS in the DLPFC led to a reduction of punishments, compared to participants with intact DLPFC excitability. Therefore the authors conclude that the DLPFC neural activity has a causal role in the willingness to punish fairness norm violators.

Furthermore, Figner's team (2010) revealed a role of the lateral prefrontal cortex (LPFC) for self-control in intertemporal choice behavior. Intertemporal choices require one to decide between receiving a smaller good (e.g. money or food, but also health benefits) in a closer future (usually immediately, but also in days or months) or a larger good in a distant future. Depending on the options an individual chooses it is then possible to measure its level of self-control: the more one prefers the distant-in-time option the higher her self-control level. Disrupting LPFC excitability by means of rTMS resulted in decreased self-control, as people chose more often the immediate smaller good over the alternative option. Taken together these studies show that brain

stimulation could influence two mechanisms strongly related to moral behavior, i.e. self-control and willingness to punish norm violators, as they are involved in social decisions where one is required to choose between a personal gain or benefiting the society (Elster, 1989, Fehr & Gächter, 2002, Fehr & Fischbacher, 2004, Crockett et al., 2010b).

Furthermore, the link between these two mechanisms and moral behavior is made more salient in a more recent study by Ruff et al. (submitted; see section 4). In this study we show that the LPFC is causally necessary to avoid altruistic punishment, inducing people to share fairly between oneself and another person when punishment for unfair behavior is allowed. More specifically, increased LPFC excitability (by means of anodal tDCS) resulted in more successful social interactions compared to decreased LPFC excitability (by means of cathodal tDCS) or natural LPFC excitability (sham stimulation). This study thus suggests that it is possible to improve moral behavior by increasing sensitivity to punishment threat, which is probably achieved as a side effect of improving self-control.

Finally, in a more recent study, Tassy and colleagues (2012), examined the effects of disrupting the right dLPFC by means of rTMS on moral judgments expressed in the context of moral dilemmas where a person is called to judge if it is morally permissible to sacrifice a small number of people (usually one) to save the lives of many more (usually 5). The evidence reported by these authors shows that compared to controls with undisrupted right PFC activity, disruption leads to a higher likelihood of making utilitarian judgments.

#### **4.8. Should we try to improve, and is it possible?**

Relying on the evidence outlined so far, this final section discusses the question of whether we should make use of the knowledge gained from empirical research on human behavior and psychology in order to improve moral practice and/or decision-making. Even if we arrive at a positive answer to this question, however, it remains unclear, how this project ought to be carried out and whether, in turn, this is possible. We will discuss the former question first and turn to the latter, the question of implementation, later.

Whether we should strive for moral improvement depends on (a) whether we believe that it is something worth striving for, (b) whether, assumed that we think it is, we should strive for it,

and (c) granted that we should, whether the methods and techniques outlined in this essay provide morally acceptable means for such a project.

(a) From a consequentialist perspective, moral improvement tends to be something worth striving for, granted that moral improvement is understood to yield overall better states of affairs. However, if one does not share such a consequentialist outlook (and many people do not), moral improvement may not appear as something worth striving for, as morality does not seem to be something we can be passionate about and desire in itself (Wolf, 1982, p. 424). In a similar vein, Williams (1981a) has argued that it is necessary for our existence to have some personal “projects”, i.e., action-guiding desires or aims which are distinct from the pure utilitarian pursuit of happiness or any other motivation derived from a moral theory.

(b) *Prima facie*, it seems odd not to strive for moral improvement if we acknowledge that it is worth striving for it. After all, it is widely assumed that if we consider something as morally good, we are motivated to act in a way to bring it about or at least not to act against it and likewise, it is assumed that if we believe we are morally required to  $\phi$ , we are motivated to  $\phi$  (internalism; Williams, 1981b, pp. 101-13). However, it is debatable whether this very assumption is correct and whether we should adopt it. On the one hand, it is highly probable that we firmly believe in something's being morally good or right and nevertheless do not act accordingly (externalism). Otherwise, the problem of *akrasia* would not even arise. On the other hand, even if a moral belief does motivate us in a certain way, this link itself may be questionable from a moral point of view. For instance, from a consequentialist perspective, it might be better if everybody acted upon certain rules laid out by some ethical framework, not upon their own moral convictions.

We will not go into a deeper discussion of motivational internalism and externalism here. A second point that can be made in this context is that, as a matter of fact, people generally strive for moral improvement or at least they claim to do so, i.e., they want to act in a more decent way, they want to become morally better individuals, they want the world to be a morally better place, etc. Three remarks will be made about this: First, the folk notions of morally good individuals, actions and states of affairs are vague and require clarification. Second, it is debatable whether people really do claim to strive for moral improvement and in which contexts and, again, what



they understand by it. Third, it may be questioned whether their claim is appropriate, i.e., whether they are in fact concerned about moral improvements or only just pretend to be (provide some possible reasons). All these questions are worth pursuing in the future.

(c) An extended debate has arisen around this question for every single method we have outlined above (e.g., for the debate on enhancement: Savulescu et al., 2011; Savulescu & Bostrom, 2009; Douglas 2008). Due to space limitations, we shall therefore only mention a few important arguments here. First, the mere possibility of moral improvement may count in favor of such a project, once it is acknowledged that moral improvement is desirable and ought to be aimed at. Furthermore, it may be viewed as an extension of methods that are already used for moral improvement at present, e.g., teaching, self-reflection, etc.

Second, and in contrast to the position just sketched, it may be doubted that any of the methods and techniques provides an acceptable way to moral improvement at all. Several reasons may be given for this position. To begin with, one may be skeptical about whether any of the approaches outlined above can really yield actual moral improvement. After all, so far only small, primarily short-term and reversible effects have been achieved. Yet, although it seems plausible that there is a limit to improvement given the constraints of the human mind and body and that moral perfection cannot be achieved, it seems doubtful that it be not possible to improve at all. The empirical evidence we have reviewed above supports this notion. Also, it may be argued that the methods for improvement are not reliable because further research is required in order to allow for their responsible application. However, it may be replied from a consequentialist perspective that such risks can be accounted for by calculating the sum of all possible outcomes each multiplied with the probability of its occurrence. For some techniques such as nudging, no morally neutral default option is available: e.g., either a country's citizens are organ donors by default or they are not, but each option invokes moral issues and there is no option outside of the moral realm. Third, a debunking argument in favor of applying the techniques and methods described could be established on the ground that all considerations speaking against such a project are merely products of a human status-quo bias. Much more could be said on each of the considerations described above. We assume that enough evidence suggests that attempts of moral improvement could be believed to be promising.

Let us now turn to the question of implementation: if we assume that we should try to achieve moral improvement, should such a project actually be carried out and if so, how? As the matter here is complex and partly speculative, we will restrict ourselves to providing a brief sketch of two issues that are relevant to this debate.

To even start considering improving moral behavior, one has to first tackle the complex philosophical problem of identifying a standard for moral improvement. This might require defining an ultimate universally accepted moral code, or agreeing on a set of general moral rules, being these either consequentialist rules or non-consequentialist ones or any other type of moral standard. Such a standard would then have to be used to gear interventions used to improve moral behavior. Whether it is in principle possible to identify such a standard, however, is highly controversial. For one thing, moral relativists hold that moral standards are relative to a culture (Wong, 1984) and thus prescribe very different behaviors. Some for instance forbid abortion while others allow it. Improving moral behavior may thus be specific for every moral community sharing the same moral standards. More profoundly, one may be skeptical about whether it is in principle possible to achieve agreement on moral questions, given that current debates about moral issues reveal both inter-cultural and intra-cultural discrepancies. For instance, from a consequentialist perspective, it may be a moral improvement to increase the number of potential organ donors, but from some religious or deontological perspectives, this would be regarded as immoral. Moreover, there is the danger of abuse by the agents or institutions in charge of implementing a process of moral improvement. Determining a prudent and trustworthy authority for this task may be extremely difficult if not impossible. Most people seem unwilling to entrust others with the care of their moral development.

Second, on a more pragmatic stance, altering moral behavior may not yield the desired improvement effects or have counterproductive side effects. For instance, promoting trustfulness may result in exploitation of trustful agents, and increasing altruistic behavior may benefit unfairly selfish individuals who could easily take advantage of altruists. In addition, the danger of a moral “lock-in” is lurking: once a process of moral improvement has begun, it may be irreversible, as the moral outlook produced by this process may prevent us from reviving lost values; mistakes may become uncorrectable.

In sum, the question of whether we should try to achieve moral improvement and whether this is possible raises a legion of extremely controversial questions. Note that the present paper itself does not mean to take a normative position on the issue of whether morality should be improved. The above points are merely meant to provide some leads for the debate.

#### **4.9. Conclusion**

The aim of this paper was to investigate why individuals often fail to judge and act in a morally decent way and what one can do about it. Investigations on morally problematic and inconsistent behavior, dominated by, e.g., cognitive biases and emotional influences, have revealed two main clusters of reasons: first, agents reason in fallacious ways, and, second, in judging or acting, they fail to account for their moral convictions. These phenomena allow for several ways of improvement. For instance, nudging may facilitate actions in accordance with moral aims, training and education may ameliorate agents' capacities for moral reasoning, pharmacological enhancement and neuro-stimulation techniques may yield improvements of both moral reflection and capacity to act morally. However, the impact and the application spectrum of all these methods have not yet been thoroughly studied, as their development is still an on-going process. An answer to the question of whether they should be implemented not only depends on future research in this field but also requires careful philosophical consideration and societal debate. We believe that these endeavors are highly relevant for a possible improvement of moral practice and therefore for the future of humanity in general.

#### **5. General Discussion**

The aim of the present thesis was twofold: on a more empirical tone it aimed at advancing the understanding of the neuro-psychological mechanisms involved in moral decision making; secondly, on a more philosophical tone, it discussed the possible consequences of having an always clearer understanding of moral decision making: such a knowledge seems to allows us to improve moral behavior, but at the same time the possibility of manipulating behavior raises deep philosophical concerns. To achieve these goals I focused in a) providing fresh evidence addressing a critical issue which has been generating numerous debates in moral philosophy and psychology: what is the involvement of emotion and cognitive processes in morality?, and b) discussing the possible repercussion of the advancement in our understanding

of how moral decisions may be manipulated: Is it technically possible to manipulate behavior? If yes, is it morally acceptable to do it, and what should guide these manipulations?

Having highlighted in a thorough review the problems and shortcomings of the existing moral psychological theories, in the first study I proposed an innovative approach to disclose the role played by emotions in moral judgments suggesting that the relation between the two can be best understood by taking into account emotions' motivational tendencies. While emotions were at the center of the first study, the second one focused on the role of cognitive control. More specifically, in study two the causal functional involvement of the LPFC in fairness judgments was tested by manipulating this brain area's excitability using tDCS. Finally, on a broader and more theoretical tone, in the last article proposed in this thesis I discussed from a novel point of view the legitimacy of using empirical data to inform moral philosophical theories, proposing that from a pragmatic point of view moral science and philosophy are required to join forces in order to achieve their ultimate goal of promoting human social coexistence.

The novel findings introduced in the first two papers unveiled some of the possible mechanisms through which emotions and cognitive processes shape our moral decisions. I chose to investigate how both types of mechanisms are involved in morality in order to strengthen the view that being morality an extremely complex function, it should not be reduced to a unique biological feature (being this feature emotions as for Haidt (2001) and Prinz (2005) or an emotionally and cognitively independent moral organ (Hauser, 2006)). In particular, with respect to emotions I suggest that these influence moral judgments depending on their motivational tendencies; while with respect to cognitive reasoning, brain activity in the LPFC determines our ability of refraining from pursuing selfish goals and conforming to the socially prescribed behavior. In other words, based on the proposed empirical findings it seems that morality is neither solely of emotional or cognitive nature, but it is actually a combination of these two. The involvement of one, the other, or both processes is modulated by the context in which one is called to make a moral decision.

With respect to the models proposed so far in the moral literature, presented thoroughly in the general introduction, the evidence provided in this work allows to exclude those models which deny an involvement of emotions in determining moral judgments, i.e. the Universal

Moral Grammar one (Mikhail, 2007; Hauser, 2006): emotions in fact have been shown to be strongly involved in several types of moral decisions. Furthermore, the models suggesting that morality is solely of emotional nature seem to be implausible as well, as the first study revealed that under certain conditions emotions do not seem to be the predominant element influencing moral judgments, and more importantly study two revealed a causal role of a cognitive process in determining moral behavior. These findings hence strongly undermine the view that emotions are the sole element having a role in morality, let alone constituting moral concepts (Prinz, 2006; Haidt, 2001). Thus the social intuitionist (Haidt, 2001) and the emotion constitutional model (Prinz, 2006) although proving accurate in grasping the relation between emotions and morality are rather incomplete as these fail to capture a relevant part of the processes involved in moral decisions.

It therefore seems that the Dual-Process model (Greene, 2001) is the most accurate, albeit least parsimonious given it holds more variables, one as it allows for both emotional and cognitive mechanisms to play a role in morality. In light of the fresh evidence proposed here, however, it is evident that this model needs to be refined: for instance based on the evidence gathered in our first study, emotions had a strong influence on both impersonal and personal moral scenarios, inducing participants to make more deontological-like or utilitarian-like moral judgments depending on the type of emotion. Hence, the notion promoted by Greene and colleagues (2004, 2002) that different types of neural mechanisms underlie different moral judgments does not seem to accurately capture how these different moral judgments are actually formed in the brain. Instead, it is more likely that different types of moral judgments could be the result of differences in the amount of information gathered through each of these mechanisms while computing this judgment. Based on the evidence provided in the two studies discussed here, I propose that the relevance of these mechanisms is strongly modulated by the context in which the moral judgment is generated. In order to better understand how these and other brain functions contribute to moral valuation future research will have to provide a mechanistic understanding of moral decision making, as it is been done in the newly defined discipline of neuroeconomics with respect to the economic decision making process (Glimcher 2009). To achieve such an understanding several studies will have to address some crucial questions such as which contextual features modulate the involvement either of emotions or of cognitive

mechanisms?; which brain areas are causally necessary to allow emotions' influence in moral judgments?; what are the neural and cognitive basis of individual differences in moral judgments?

The possibility of achieving this level of knowledge on moral decision making opens a broad range of empirical and philosophical questions, some of which have been addressed in the third paper proposed in this dissertation: on a practical level, do we have the means to manipulate the moral decision processes? More specifically, in this paper we provided examples of known situations in which people are likely to make mistakes in moral reasoning, leading to inappropriate or incoherent moral behavior, exploring some of the potential reasons justifying these mistakes and behaviors. Furthermore, we addressed some of the pioneering studies attempting to influence moral behavior through different interventions, as for instance pharmacological manipulations, training programs or brain stimulation. Through the discussion of the above mentioned issues, from a more philosophical point of view this paper finally discussed some of the moral dilemmas we necessarily have to face if one considers manipulating moral behavior. Assuming that soon we might be in possess of the means to manipulate moral behavior, is it morally appropriate to manipulate morality? Which moral rules have to be taken into consideration to guide these manipulations? While these questions are of purely normative/philosophical nature, providing answers will require a deep knowledge of both the practical and ethical implications of intervening on moral behavior. For this reasons it is claimed in this dissertation that philosophers and scientists will be required to join forces.

### **5.1.Limitations of this thesis**

While trying to be as thorough as possible in all its parts, this thesis has inevitably some limitations. A very important part of this thesis focused on the role of emotions for moral judgments. Being focused on the main models describing morality and moral decision making, in this thesis I did not discuss in depth all the properties of emotions and how these different properties might help better understand the different mechanisms through which these might influence moral judgments, including the one I investigated in Study I, i.e. motivational tendencies. The same hold for cognitive mechanisms: the way in which I have referred to these throughout my thesis did not take into account a lot of several minute aspects characterizing

them. On a more methodological level, this thesis's strongest limitation is possibly resulting by having employed very different experimental paradigms in Study I and II. The first one required people to imagine complex moral situations, while the second study demanded people to allocate money between themselves and a stranger. Therefore, although both studies tested people's moral decision making, a direct comparison between the results of these two studies is somehow hindered.

## **5.2. Conclusion**

In this dissertation I critically analyzed the existing models describing moral decision making. Through this critical analysis were identified some of the most relevant sources of controversy between the scholars studying moral decision making, most notably: the role of emotions and cognitive mechanisms in morality. Two studies were performed precisely to address this controversy, the results revealing that both have an important role in shaping moral judgments and actions, and that their involvement is strongly influenced by circumstances. In light of the new evidence, it seems that moral descriptive models have to grant a role to both emotions and cognitive mechanisms in order accurately capture how morality is processed. We suggest that a satisfactory descriptive model of moral decision making could be achieved adopting a mechanistic approach.

## 6. References

- Allen, R.E. (2006). *Plato: The Republic*. New Haven: Yale University Press.
- Althof, W., and Berkowitz, M. W. (2006). Moral education and character education: their relationship and roles in citizenship education. *Journal of Moral Education*, 35, 495-518.
- Anderson, S. W., Barrash, J., Bechara, A., Tranel, D. (2006). Impairments of emotion and real-world complex behavior following childhood- or adult-onset damage to ventromedial prefrontal cortex. *Journal of the International Neuropsychological Society*, 12, 224-235.
- Andreoni, J., Harbaugh, W., and Vesterlund, L. (2003). The carrot or the stick: rewards, punishments, and cooperation. *American Economic Review*, 93, 893–902.
- Andreoni, J. (1990). "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow
- Aron, A.R., Robbins, T.W., and Poldrack, R.A. (2004). Inhibition and the right inferior frontal cortex. *Trends Cogn. Sci.* 8, 170–177.
- Giving". *Economic Journal* 100 (401): 464–477
- Apperly, I. A. (2008) Beyond Simulation-Theory and Theory-Theory: why social cognitive neuroscience should use its own concepts to study “theory of mind”. *Cognition*, 107, 266-283.
- Austin, J. (1961). A plea for excuses. In his *Philosophical Papers*. Oxford: Clarendon.
- Balleine, B. W., Liljeholm, M., Ostlund, S. B. (2009). The integrative function of the basal ganglia in instrumental conditioning. *Behavioural Brain Research*, 199, 43-52.
- Baumeister, R. F., Heatherton, T. F., Tice, D. M. (1994). *Losing control: How and why people fail at self-regulation*. San Diego, CA: Academic Press.
- Bear, M. F., Connors, B. W., & Paradiso, M. A. (Eds.). (2001). *Neuroscience: Exploring the brain* (2nd ed.). Baltimore: Lippincott Williams & Wilkins.
- Bebeau, M. J., Rest, J. R., Narvaez, D. (1999). Beyond the promise: A perspective on research in moral education. *Educational Researcher*, 28, 18-26.



- Bentham, J. (1996/1789), *An Introduction to the Principles of Morals and Legislation*, J.H. Burns and H.L.A. Hart (eds), Oxford: Clarendon Press.
- Berkowitz, L. (2003). Affect, aggression, and antisocial behavior. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 804-823). New York, NY: Oxford University Press.
- Blair, R. J. (2007). The amygdala and ventromedial prefrontal cortex in morality and psychopathy. *Trends in Cognitive Sciences*, 11, 387-392.
- Blair, R. J. R., Mitchell, D., & Blair, K. (2005). *The psychopath: Emotion and the brain*. New York, NY: Wiley-Blackwell.
- Blair, R.J.R., and Mitchell, D.G.V., Richell, R.A., Kelly, S., Leonard, A., Newman, C. and Scott, S.K. (2002). Turning a Deaf Ear to Fear: Impaired Recognition of Vocal Affect in Psychopathic Individuals. *Journal of Abnormal Psychology*, 111, 682–686.
- Blair, R.J.R. (1995). A Cognitive Developmental Approach to Morality: Investigating the Psychopath. *Cognition*. 57, 1-29.
- Blasi, A. (1980). Bridging moral cognition and moral action: A critical review of the literature. *Psychological Bulletin*, 88, 593-63.
- Boik, R. J. (1981). A priori tests in repeated measures designs: Effects of nonsphericity. *Psychometrika*, 46, 241-255.
- Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., and Cohen, J.D. (2001). Conflict monitoring and cognitive control. *Psychol. Rev.* 108, 624–652.
- Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A. R., Langner, R., Eickhoff, S. B. (2012). Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure & Function* (in press).
- Cato, M.A., Crosson, B., Gokcay, D., Soltysik, D., Wierenga, C., Gopinath, K., Himes, N., Belanger, H., Bauer, R.M., Fischler, I.S., Gonzalez- Rothi, L., Briggs, R.W., (2004) Processing

words with emotional connotation: an fMRI study of time course and laterality in rostral frontal and retrosplenial cortices. *J. Cogn. Neurosci.* 16 (2), 167–177.

Chomsky, N. (2000) *New Horizons in the Study of Language and Mind*. Cambridge, England: Cambridge University Press.

Chomsky, N. (1988) *Generative Grammar: Its Basis, Development and Prospects*. Studies in English Linguistics and Literature, Special Issue, Kyoto: Kyoto University of Foreign Studies.

Chomsky, N. (1986) *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger Publishers.

Ciaramelli, E., Muccioli, M., Làdavas, E., di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, 2, 84-92.

Cima, M., Tonnaer, F., & Hauser, M. D. (2010). Psychopaths know right from wrong but don't care. *Social Cognitive Affective Neuroscience*, 5, 59-67.

Cohen, M. X., Krohn-Grimberghe, A., Elger, C. E., Weber, B. (2007). Dopamine gene predicts the brain's response to dopaminergic drug. *European Journal of Neuroscience*, 26, 3652-3660.

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J. & Smith, E. E. (1997) Temporal dynamics of brain activation during a working memory task. *Nature* 386, 604 – 608. doi:10.1038/386604a0.

Cools, R., D'Esposito, M (2011). Inverted-U-shaped dopamine actions on human working memory and cognitive control. *Biological Psychiatry*, 69, e113-e125.

Courtney AE, Maxwell AP. (2009) The challenge of doing what is right in renal transplantation: balancing equity and utility. *Nephron, Clin. Pract.*, 111 (1).

Crockett, M. J., Clark, L., Hauser, M. D., Robbins, T. W. (2010a). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences*, 107, 17433-17438.

- Crockett, M. J., Clark, L., Lieberman, M. D., Tabibnia, G., Robbins, T. W. (2010b). Impulsive choice and altruistic punishment are correlated and increase in tandem with serotonin depletion. *Emotion*. 10(6):855-62.
- Cushman, F., Young, L., Greene, J. (2010). Our multi-system moral psychology: Towards a consensus view. In J. Doris, G. Harman, S. Nichols, J. Prinz, W. Sinnott-Armstrong, S. Stich. (eds.), *The Oxford Handbook of Moral Psychology*. Oxford University Press.
- Damasio AR, Grabowski TJ, Bechara A, Damasio H, Ponto LL, Parvizi J, Hichwa RD (2000) Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nat Neurosci* 3:1049–1056.
- Darley, J. M., Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, 8, 377–383.
- Decety, J., Lamm, C. (2007) The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. *Neuroscientist*, 13(6), 580-593.
- Delgado, M. R., Frank, R. H., Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, 8, 1611-1618.
- Douglas, T. (2008). Moral enhancement. *Journal of Applied Philosophy* 25(3):228-45.
- Driver, J., Blankenburg, F., Bestmann, S., Vanduffel, W. & Ruff, C.C. (2009) Concurrent brain-stimulation and neuroimaging for studies of cognition. *Trends in Cognitive Sciences* 13, 319-327.
- Dupoux, E., & Jacob, P. (2007) Universal moral grammar: a critical appraisal. *Trends in cognitive sciences*, 11(9), 373-378.
- Dwyer, S. (1999). Moral competence. In K. Murasugi and R. Stainton (Eds.), *Philosophy and Linguistics*. Linguistics. Westview Press.
- Ekman, P. (1972). *Emotions in the Human Face*. New York: Pergamon Press.

- Elster, J. (1989). *The Cement of Society - A Study of Social Order* (Cambridge: Cambridge University Press).
- Eskine, K. J., Kacinik, N. A., Prinz, J. J. (2011). A bad taste in the mouth: gustatory disgust influences moral judgment. *Psychological Science* 22(3):295-9.
- Farber, P. L. (1994). *The Temptations of Evolutionary Ethics*. Berkeley: University of California Press.
- Fecteau, S. et al. (2007). Diminishing risk-taking behavior by modulating activity in the prefrontal cortex: a direct current stimulation study. *Journal of Neuroscience*, 27, 12500-12505.
- Fehr, E., Fischbacher, U. (2004). Third-party punishment and social norms. *Evolutionary Human Behavior*, 25, 63–87.
- Fehr, E., Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415, 137–140.
- Fehr, E., Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114, 817-868.
- Figner, B., Knoch, D., Johnson, E., Krosch, A., Lisanby, S., Fehr, E., Weber, E.U. (2010) Lateral prefrontal cortex and self-control in intertemporal choice. *Nature Neuroscience*, 13, 538–539.
- Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, 71, 397–404.
- Fitzgerald, P. B., Fountain, S., Daskalakis, J. (2006). A comprehensive review of the effects of rTMS on motor cortical excitability and inhibition. *Clinical Neurophysiology*, 117, 2584–2596.
- Foot, P. R. (1978) *Virtues and Vices and Other Essays in Moral Philosophy*. Berkeley: University of California Press; Oxford: Blackwell.
- Foot, P. R. (1967) *An Existentialist Ethics*. Hazel E. Barnes Knopf, 462 pp.
- Forbes, C. E., Grafman J. (2010). The role of the human prefrontal cortex in social cognition and moral judgment. *Annual Review of Neuroscience*, 33, 299-324.

Forgas, J. P. (2003). Affective influences on attitudes and judgments. In R. J. Davidson, K. J. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 596-618). New York, NY: Oxford University Press.

Glimcher, P.W. (2009). Introduction: A Brief History of Neuroeconomics. In: Glimcher, P.W., Camerer, C.F., Fehr, E., and Poldrack, R.A. (eds.) *Neuroeconomics: Decision Making and the Brain*. New York: Academic Press, pp.1-12.

Green L, Myerson J. (2004) A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*, 130, 769–792.

Green L, Myerson J, Holt DD, Slevin JR, Estle SJ. (2004) Discounting of delayed food rewards in pigeons and rats: Is there a magnitude effect? *Journal of Experimental Animal Behavior*, 81, 39–50.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, 111, 364-371.

Greene, J., Morelli, S., Lowenberg, K., Nystrom, L., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, 107, 1144-1154.

Greene, J. D. (2007a). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, 11(8), 322–323.

Greene, J. D. (2007b). The secret joke of Kant's soul. In: W. Sinnott-Armstrong, ed. *Moral Psychology*, Vol. 3. Cambridge (MA): MIT Press, pp. 35-79.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44, 389-400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293, 2105-2108.

Guth, W., Schmittberger, R., and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *J. Econ. Behav. Organ.* 3, 367–388.

- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108, 814-834.
- Haidt, J., & Joseph, C. (2004). Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. *Daedalus*, pp. 55-66, Special issue on human nature.
- Harlé, K. M., & Sanfey, A. G. (2010). Effects of approach and withdrawal motivation on interactive economic decisions. *Cognition & Emotion*, doi:10.1080/02699930903510220.
- Hamilton, W. D., 1964, 'The genetical evolution of social behaviour I and II', *Journal of Theoretical Biology*, 7, 1-32
- Harenski, C. L., Hamann, S. (2006). Neural correlates of regulating negative emotions related to moral violations. *Neuroimage*, 30, 313-324.
- Harmon-Jones, E. (2004). Contributions from research on anger and cognitive dissonance to understanding the motivational functions of asymmetrical frontal brain activity. *Biological Psychology*, 67, 51-76.
- Harmon-Jones, E., & Sigelman, J. (2001). State anger and prefrontal brain activity: Evidence that insult-related relative left prefrontal activation is associated with experienced anger and aggression. *Journal of Personality and Social Psychology*, 80, 797-803.
- Harrison, N. A., Wilson, C. E., & Critchley, H. D. (2007) Processing of observed pupil size modulates perception of sadness and predicts empathy. *Emotion*, 7, 724-729.
- Hauser, M.D., L. Young & F.A. Cushman (2008) Reviving Rawls's linguistic analogy. In *Moral Psychology and Biology* ed. W. Sinnott-Armstrong, Oxford University Press, NY.
- Hauser, M. (2006) *Moral Minds: How Nature Designed a Universal Sense of Right and Wrong* (Harper Collins/Ecco, NY).
- Hebb, D.O. (1949) *The Organization of Behavior*, Wiley: New York.

- Heekeren, H. R., Wartenburger, I., Schmidt, H., Prehn, K., Schwintowski, H. P., Villringer, A. (2005). Influence of bodily harm on neural correlates of semantic and moral decision-making. *Neuroimage*, 24, 887-897.
- Heekeren, H. R., Wartenburger, I., Schmidt, H., Schwintowski, H. P., Villringer, A. (2003). An fMRI study of simple ethical decision-making. *Neuroreport*, 14, 1215-1219.
- Held, V. (2002). Moral Subjects: The natural and the normative. *Proceedings and Addresses of the American Philosophical Association*, 76(2), 7-24.
- Huebner, B., Dwyer, S., & Hauser, M. D. (2009). The role of emotion in moral psychology. *Trends in Cognitive Science*, 13, 1-6.
- Hume, D. (1960). *An enquiry concerning the principles of morals*. La Salle, IL: Open Court. (Original work published 1777).
- Illes, J., Sahakian, B. J. (eds., 2011). *The Oxford handbook of neuroethics*. Oxford: University Press.
- Iyer, M.B. et al. (2005). Safety and cognitive effect of frontal DC brain polarization in healthy individuals. *Neurology*, 64, 872-875.
- Jaeggi, S. M., Buschkuhl, M., Jonides, J., Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, 105, 6829-6833.
- Jaeggi, S. M., Buschkuhl, M., Jonides, J., Shah, P. (2011). Short- and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences*, 108, 10081-10086.
- Jones, R. W., Morris, K., Nutt, D. (2005). *Cognition enhancers. Foresight brain science addiction and drugs project*. London: Office of Science and Technology.
- Kahneman, D., Wakker, P. P., Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *Quarterly Journal of Economics*, 112, 375-405.

- Kahneman, D., Knetsch, J.L., and Thaler, R. (1986). Fairness as a constraint on profit seeking - entitlements in the market. *Am. Econ. Rev.* 76, 728–741.
- Kalis, A., Mojzisch, A., Schweizer, T. S., Kaiser, S. (2008). Weakness of will, akrasia, and the neuropsychiatry of decision making: an interdisciplinary perspective. *Cognitive, Affective and Behavioral Neuroscience*, 8, 402-417.
- Kamm, F. (2000) Nonconsequentialism, in Hugh LaFollette *The Blackwell Guide to Ethical Theory*, Oxford, pp. 205–26.
- Kant, I. (1959). *Foundations of the metaphysics of morals* (L. W. Beck, Trans.). Indianapolis, IN: Bobbs-Merrill. (Original work published 1785).
- King, P. M., Mayhew, M. J. (2002). Moral judgement development in higher education: Insights from the defining issues test. *Journal of Moral Education*, 31, 247-270.
- Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., Gillberg, C. G., Forssberg, H., Westerberg, H. (2005). Computerized training of working memory in children with ADHD--a randomized, controlled trial. *Journal of the American Academy of Child and Adolescent Psychiatry*, 44, 177-186.
- Knoch, D., Nitsche, M. A., Fischbacher, U., Eisenegger, C., Pascual-Leone, A., Fehr, E. (2008). Studying the neurobiology of social interaction with transcranial direct current stimulation--the example of punishing unfairness. *Cerebral Cortex*, 18, 1987-1990.
- Knutson, K.M., Krueger, F., Koenigs, M., Hawley, A., Escobedo, J.R., Vasudeva, V., Adolphs, R., and Grafman, J. (2010). Behavioral norms for condensed moral vignettes. *SCAN*, 5(4), 378-384.
- Knutson, B., Taylor, J., Kaufman, M., Peterson, R., and Glover, G. (2005). Distributed neural representation of expected value. *J. Neurosci.* 25, 4806–4812
- Kocha, A. K., Normann, HT., (2008) Giving in Dictator Games: Regard for Others or Regard by Others? *Souther Economic Journal*, 75(1),223-233



- Koenigs, M., Young, L., Adolph, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, 446, 908-911.
- Kohlberg, L; T. Lickona, ed. (1976). "Moral stages and moralization: The cognitive-developmental approach". *Moral Development and Behavior: Theory, Research and Social Issues*. Holt, NY: Rinehart and Winston.
- Korsgaard, C. M. (1996). *The Sources of Normativity*. Cambridge: Cambridge University Press.
- Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., Fehr, E. (2005). Oxytocin increases trust in humans. *Nature*, 435, 673-676.
- Kosslyn, S. M., Shin, L. M., Thompson, W. L., McNally, R. J., Rauch, S. L., Pitman, R. K., and Alpert, N. M. (1996). Neural effects of visualizing and perceiving aversive stimuli: A PET investigation. *NeuroReport*, 7, 1569-1576. E. M. Reiman et al., *Am. J. Psychiatry* 154, 918 (1997).
- Kringelbach, M.L. (2005). The human orbitofrontal cortex: linking reward to hedonic experience. *Nat. Rev. Neurosci.* 6, 691–702.
- Lamm, C., & Singer, T. (2010). The role of anterior insular cortex in social emotions. *Brain Struct Funct*, 214(5-6), 579-591.
- Lane R, Chua P, Dolan R (1999) Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures. *Neuropsychologia* 37:989–997.
- Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). Motivated attention: Affect, activation and action. In P. J. Lang, R. F. Simons, & M. T. Balaban (Eds.), *Attention and orienting: Sensory and motivational processes* (pp. 97-135). Hillsdale, NJ: Erlbaum.
- Lawrence, A. D., and Calder, A. J. (2004). Homologizing Human Emotions. In D. Evans and P. Cruse (Eds.), *Emotions, Evolution and Rationality* (pp. 15-47). Oxford: Oxford University Press.

- Lazarus, R. S. (1999). The cognition-emotion debate: a bit of history. In T. Dalgleish & M. J. Power (Eds.), *Handbook of Emotion and Cognition* (pp. 3-19). Chichester, New York: John Wiley and sons.
- Leiberg, S., Klimecki, O., Singer, T. (2011). Short-term compassion training increases prosocial behavior in a newly developed prosocial game. *Public Library of Science ONE*, 6, e17798.
- Li, S. C., Schmiedek, F., Huxhold, O., Röcke, C., Smith, J., Lindenberger, U. (2008). Working memory plasticity in old age: practice gain, transfer, and maintenance. *Psychology and Aging*, 23, 731-742.
- Mackie, J. L. (1980). *Hume's Moral Theory*. London: Routledge and Kegan Paul Ltd.
- MacLeod, C. M. (1991) Half a century of research on the Stroop effect: an integrative review. *Psychol. Bull.* 109 (2), 163-203.
- Manning, R., Levine, M. & Collins, A. (2007). The Kitty Genovese murder and the social psychology of helping: The parable of the 38 witnesses. *American Psychologist*, 62(6), 555-562.
- Maddock, R.J. (1999) The retrosplenial cortex and emotion: new insights from functional neuroimaging of the human brain. *Trends Neurosci.* 22, 310-6
- Masten, C. L., Morelli, S. A., Eisenberger, N. I. (2011). An fMRI investigation of empathy for 'social pain' and subsequent prosocial behavior. *Neuroimage*, 55, 381-388.
- McClure, S. M., Ericson, K.M., Laibson, D. I., Loewenstein, G. & Cohen, J. D. (2007) Time discounting for primary rewards. *Journal of Neuroscience*, 27, 5796-5804.
- McClure, S. M., Laibson, D. I., Loewenstein, G., Cohen, J. D. (2004) Separate neural systems value immediate and delayed monetary rewards. *Science*, 306, 503–507.
- Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, 11, 143-152.

- Mikhail, J. (2000). Rawls' linguistic analogy: A study of the "generative grammar" model of moral theory described by John Rawls in a theory of justice. Unpublished doctoral dissertation, Cornell University, Ithaca, NY.
- Miller, E.K., and Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24, 167–202.
- Mitchell, J. P. (2005) The false dichotomy between simulation and theory-theory: the argument's error. *Trends in Cognitive Science*, 9 (8), 363-364.
- Moll, J., De Oliveira-Souza, R., Zahn, R. (2008) The neural basis of moral cognition: sentiments, concepts, and values. *Annals of the New York Academy of Sciences*, 1124, 161-180.
- Moll, J. & de Oliveira-Souza, R. (2007) Moral judgments, emotions and the utilitarian brain. *Trends in cognitives sciences*, 11(8), 319-321.
- Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., Grafman, J. (2005) Opinion: the neural basis of human moral cognition. *Nature reviews. Neuroscience*. 6(10), 799-809.
- Moll, J., De Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourao-Miranda, J., Andreiuolo, P. A., & Pessoa, L. (2002). The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions. *Journal of Neuroscience*, 22, 2730-2736.
- Moore, G. E. (1903), *Principia Ethica*. Cambridge: Cambridge University Press.
- Moretto, G., Làdavas, E., Mattioli, F., di Pellegrino, G. (2010). A psychophysiological investigation of moral judgment after ventromedial prefrontal damage. *Journal of Cognitive Neuroscience*, 22, 1888-1899.
- Müller-Lyer, F. C. (1889). Optische Urteilstäuschungen. *Archiv für Physiologie Suppl.* 263–270.
- Nietzsche, F. (1966). *Beyond Good and Evil*, trans. W. Kaufmann, New York: Vintage.

- O'Doherty, J., Kringelbach, M.L., Rolls, E.T., Hornak, J., and Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nat. Neurosci.* 4, 95–102.
- Olsson, A., Nearing, K.I., Phelps, E. A. (2007). Learning fears by observing others: the neural systems of social fear transmission. *Social Cognitive and Affective Neuroscience*, 2, 3-11.
- Piaget, J. (1932). *The Moral Judgment of the Child*. London: Kegan Paul, Trench, Trubner and Co.
- Pigden, C. R. (1991). Naturalism. In P. Singer (Ed.), *A Companion to Ethics* (pp. 421-431): Blackwell.
- Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron*, 72, 692-697.
- Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, 10, 59-63.
- Prinz, J. (2007). *The emotional construction of morals*. New York: Oxford University Press.
- Prinz, J. (2006). The emotional basis of moral judgments. *Philosophical Explorations*, 9, 29-43.
- Power, M. & Dalgleish, T. (2008) *Cognition and Emotion: From Order to Disorder* (second edition). Hove, U.K.: Taylor Francis.
- Raine A., Yang, Y. (2006). Neural foundations to moral reasoning and antisocial behavior. *Social Cognitive and Affective Neuroscience*, 1, 203-213.
- Rawls, J. (1971). *A Theory of Justice*. Cambridge: Harvard University Press.
- Roger, C. (2000), *Aristotle: Nichomachean Ethics*. Cambridge University Press.
- Rosenthal, A. M. (1964). *Thirty-eight witnesses*. New York: McGraw-Hill.

- Rozin, P. et al. (1999) The moral-emotion triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral ethics (community, autonomy, divinity). *J. Pers. Soc. Psychol.* 76, 574–586.
- Rueda, M. R., Rothbart, M. K., McCandliss, B. D., Saccomanno, L., Posner, M. I. (2005). Training, maturation, and genetic influences on the development of executive attention. *Proceedings of the National Academy of Sciences*, 102, 14931-14936.
- Ruff, C. C., Ugazio, G., Fehr, E. (submitted). A causal role for the lateral prefrontal cortex in social norm compliance.
- Ruse, M. (1995). *Evolutionary Naturalism: Selected essays*. London: Routledge.
- Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., Cohen, J. D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, 300, 1755-1758.
- Savulescu, J., ter Meulen, R., Kahane, G. (eds.). (2011). *Enhancing Human Capacities*. Oxford: Wiley Blackwell.
- Savulescu, J., Bostrom, N. (2009). *Human Enhancement*. Oxford: Oxford University Press.
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in “theory of mind.” *Neuroimage*, 19, 1835-1842.
- Scanlon, T. (2008). *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge: Harvard University Press.
- Scheffler, S. (1988). *Consequentialism and Its Critics*. Oxford: Oxford University Press.
- Schlaefli, A., Rest, J. R., Thoma, S. J. (1985). Does moral education improve moral judgment? A meta-analysis of intervention studies using the defining issues test. *Review of Educational Research*, 55, 319-352.
- Schnall, S., Haidt, J., & Clore, G. L. (2008a). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, 34, 1096-1109.

- Schnall, S., Benton, J., Harvey, S. (2008b) With a Clean Conscience Cleanliness Reduces the Severity of Moral Judgments. *Psychological Science*, 19, 2, 1219-1222.
- Schultz, W. (2007). Multiple dopamine functions at different time courses. *Annual Review of Neuroscience*, 30, 259-288.
- Shenhav, A., Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, 67, 667-677.
- Singer, P. (1974). *Practical Ethics*. Second Edition. Cambridge: Cambridge University Press.
- Singer, T. (2006). The neuronal basis and ontogeny of empathy and mind reading: Review of literature and implications for future research. *Neuroscience & Biobehavioral Reviews*, 30, 855-863.
- Singer, T., & Lamm, C. (2009). The social neuroscience of empathy. *The Year in Cognitive Neuroscience 2009: Annals of the New York Academy of Sciences*, 1156, 81-96.
- Smith, A. (2010) *Theory of Moral Sentiments* (London: Penguin Classics, original work published in 1759).
- Smith, E. E., Jonides, J., (1997) Working memory: A view from neuroimaging. *Cognit. Psychol.*, 33, pp. 5-42.
- Sober, E. and Wilson D.S., 1998, *Unto others: The evolution and psychology of unselfish behavior*, Cambridge MA: Harvard University Press.
- Spielberg, J. M., Stewart, J. L., Levin, R. L., Miller, G. A., & Heller, W. (2008) Prefrontal cortex, emotion, and approach/withdrawal motivation. *Social and Personality Psychology Compass*, 2, 135-153.
- Spitzer, M., Fischbacher, U., Herrnberger, B., Gron, G., & Fehr, E. (2007). The neural signature of social norm compliance. *Neuron* 56, 185–196.
- Strayhorn, J. M. (2002). Self-control: Theory and research. *Journal of the American Academy of Child & Adolescent Psychiatry*, 41, 7-16.

Tassy, S., Oullier, O., Duclos, Y., Coulon, O., Mancini, J., Deruelle, C., Attarian, S., Felician, O., Wicker B. (2012) Disrupting the right prefrontal cortex alters moral judgement, *SCAN*, 7 (3), 282-288.

Thomas Aquinas, *Summa Theologiae* (IIa- IIae Question 64 article 7).

Thomas, B. C., Croft, K. E., Tranel, D. (2011). Harming kin to save strangers: further evidence for abnormally utilitarian moral judgments after ventromedial prefrontal damage. *Journal of Cognitive Neuroscience*, 23, 2186-2196.

Thomson, J. J. (1990). *The Realm of Rights*. Cambridge, MA: Harvard University Press.

Thomson, J. J. (1986). *Rights, restitution and risk*. Cambridge, MA: Harvard University Press.

Tobler, P. N., Kalis, A., Kalenscher T (2008). The role of moral utility in decision making: an interdisciplinary framework. *Cognitive Affective and Behavioral Neuroscience*, 8, 390-401.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology* 46, 35-57.

Tversky, A., Kahneman, D. (1981) The Framing of decisions and the psychology of choice. *Science* 211, 453–458.

Ugazio, G., Heinzelmann, N., Tobler, P.N. (Under Review) Pragmatic implications of empirically studying moral decision-making.

Ugazio, G., Lamm, C. and Singer, T. (2011, Epub ahead of Print) The Role of Emotions for Moral Judgments Depends on the Type of Emotion and Moral Scenario.

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, 17, 476-477.

von Neumann, J., Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. 2nd ed. Princeton, NJ: Princeton Univ. Press.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, 16, 780-784.

Wong, D.B., 1984, *Moral Relativity*, Berkeley CA: University of California Press.

Woodward, J., Allman, J. (2007). Moral intuition: its neural substrates and normative significance. *Journal of Physiology-Paris*, 101, 179-202.

Young, L., Dungan, J. (2012). Where in the brain is morality? Everywhere and maybe nowhere. *Social Neuroscience*, 7, 1-10.

Young, L., Camprodon, J., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporo-parietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgment. *Proceedings of the National Academy of Sciences of the USA*, 107, 6753-6758.

Young, L., Cushman, F., Hauser, M. D., & Saxe, R. (2006). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the USA*, 104, 8235-8240.

Young, L., & Koenigs, M. (2007). Investigating emotion in moral cognition: a review of evidence from functional neuroimaging and neuropsychology. *British Medical Bulletin*, 84, 67-79.

Zak, P. J., Stanton, A. A., Ahmadi, S. (2007). Oxytocin increases generosity in humans. *Public Library of Science ONE*, 2, e1128.



## Dr. des. Giuseppe Ugazio

Nationality: Italian

Date of Birth: 07.08.1986

Address: Bluemlisalpstrasse 10, 8006 Zurich

Email: [giuseppe.ugazio@econ.uzh.ch](mailto:giuseppe.ugazio@econ.uzh.ch)

Phone: +41 44 634 52 44

### Summary

- Primary interested in the understanding of the psychological mechanisms on which moral decisions rely and their neural correlates.
- Primary interested in the nature of Human Rights, researching innovative ways to promote and enforce these having a better understanding of the psychological processes which motivate either their implementation or their violation.
- Very interested in the emergence and dynamics of social cooperation, with particular emphasis on the relevance of moral and social norms;
- Very interested in political philosophy, with emphasis on the implementation of social choice methods in democracy;
- Possesses a very good knowledge of epistemology and philosophy of social sciences;
- Possesses a good knowledge of the traditional and more recent experimental methodologies.

### Education

**University of Zurich, Department of Economics, Zurich CH**

**PhD**, Philosophical Doctoral Program in Neuroeconomics at the Department of Economics

*October  
2011 -  
present*

Research interests mainly focused on identifying the neural substrates of individual differences in value-based moral decision making. In particular we seek to understand:

- Which functional and structural brain features can explain individual differences in moral choices?;
- To which extent do moral decisions draw on similar types of processes as other types of decisions (e.g., economic);
- Which brain regions and processes are causally necessary for moral behaviors

*September  
2008-*

**University of Zurich, Department of Economics, Zurich CH**

**PhD**, Philosophical Doctoral Program in Philosophy in the University Priority Program "Foundations of Human Social Behavior".

*November  
2012*

Research interests mainly focused on the nature of moral judgments and morality, the emotional mechanisms involved in moral decisions and the enforcement of moral norms. These issues are approached combining empirical experiments and philosophical theoretical framing. In the empirical part of our research we study whether:

- 
- Emotions selectively influence our perception of moral good and wrong, thus our moral judgments, in certain but not all situations;
- The Lateral Prefrontal Cortex plays a crucial functional and causal role in the compliance of norms, such as fairness;

In the philosophical domain we propose:

- A pragmatic approach to morality from which addressing the long-lasting *is/ought* debate.

**London School of Economics, London UK**

**MSc, Master of Sciences in Philosophy of the Social Sciences.**

*October  
2007-  
October  
2008*

Emphasis on the understanding of Individual and Organizational Behavior. Emphasis on the Neuro-Biological Foundations of Ethics.:

- Philosophy, Morals and Politics (Philosophy Department);
- Philosophy of the Social Sciences (Philosophy Department);
- Social Choice and Democracy (Government Department);
- Social and Organizational Behavior (Social Psychology Department).

**London School of Economics, London UK**

**Certificate in International Human Rights Law and Practice**, accredited by both the Law Society and the Bar Council for 20 CPD points. Main topics covered :

*July 2007-  
December  
2007*

- The understanding of internationally recognised human rights instruments and standards;
- How human rights standards work and shows how they can be applied in practice;
- How human rights standards influence the development of law and policy.

**Carnegie Mellon University, Pittsburgh/Pennsylvania, USA,**

**Summer School of Logic and Formal Epistemology**

Scholarship to participate in the Carnegie Mellon Summer School in Logic and Formal Epistemology.

*June 12*

*June 30*

*2006*

- This was an intensive three-week workshop with daily morning and afternoon sessions. Each week of workshop featured lectures devoted to a single topic:
  1. Week 1: Causal Inference (David Danks)
  2. Week 2: Foundations of Computability (Wilfried Sieg)
  3. Week 3: Philosophical Logic (Horacio Arlo-Costa)
- Workshops activities included individual and group assignments

**Università “Vita-Salute San Raffaele”, Milan I**

Undergraduate course in Philosophy,

*October  
2004*

*July 2007*

- Graduated with an overall mark of 106/110 and a GPA of 28,6/30 (Last two years GPA 29,18/30). Emphasis on the Philosophy of Sciences and Political Sciences.
- Intensive research work in the main Philosophical and Social Scientific Journals to write a final dissertation and various working papers.

**European School, Varese, I**

**European Baccalaureate**

*July 2004*

Emphasis on Philosophy, Biology, History during the last years, as well as English and Spanish as foreign languages. Also a basic knowledge of Latin.

<b>Professional Experiences</b>	<b>University of Zurich, Institute for Empirical Research in Economics, Zurich CH</b>	
	<b>Research Assistant</b>	
	<ul style="list-style-type: none"> <li>• Extensive review of the publications in the most relevant journals in the areas of Philosophy, Psychology, Neurosciences and Behavioral Economics.</li> <li>• Development of experiments to test empirically innovative and existing hypothesis on moral behavior</li> </ul>	October 2008, Present
	<b>A.I.S.P.O (Italian Association for the Solidarity Between Populations), Milan I</b>	
	<ul style="list-style-type: none"> <li>• Voluntary Working mainly in the fund raising and human resources areas with A.I.S.P.O., an international NGO operating in the health assistance field, providing high quality medical services to less developed countries.</li> </ul>	
	<b>Geneva International Model United Nations 2007 Geneva CH</b>	
	<ul style="list-style-type: none"> <li>• I have been delegate of Switzerland in the Human Rights Council</li> </ul>	March 2007
	<b>European Commission Representation in Milan, Milan, I</b>	
	<b>Intern</b>	October December 2006
	<ul style="list-style-type: none"> <li>• I've been working in a team of four students monitoring the daily press</li> <li>• I wrote reports and trends in the weekly newsletter of the Italian representation of the EU Commission</li> <li>• I participated to the daily video-conference with the spoke persons of the European Commissioners</li> <li>• I've been one of the responsible for the organization of the event and I have written reports on the discussion in a workshop of the Empower, European Civil Society Forum at Bergamo 9<sup>th</sup>-10<sup>th</sup> November 2006</li> </ul>	
	<b>Veterinary Department of Dr. Giorgio Zappellini, Cittiglio, VA, I</b>	July 2002
	<ul style="list-style-type: none"> <li>• I've done a stage which lasted one month;</li> <li>• I've worked as an assistant in the department.</li> </ul>	
	<b>Scuola Europea Varese, Varese, I</b>	
	<ul style="list-style-type: none"> <li>• I've been delegate of the European School of Varese at the Conseil Supérieur des Elèves representing my school during meetings with the other European School delegates discussing pupil's committees work in each institution.</li> </ul>	September- July 2004
<b>Spoken Languages</b>	<ul style="list-style-type: none"> <li>• ITALIAN: Mother Tongue</li> <li>• ENGLISH: WRITTEN: Excellent; SPOKEN: Excellent.</li> <li>• SPANISH: WRITTEN: Very Good; SPOKEN: Excellent.</li> <li>• GERMAN: WRITTEN: Basic; SPOKEN: Basic</li> </ul>	

<b>Publications</b>	<ul style="list-style-type: none"> <li>• <b>Ugazio, G.</b>, Lamm, C. and Singer, T. (2011) The Role of Emotions for Moral Judgments Depends on the Type of Emotion and Moral Scenario. <i>Emotion</i></li> <li>• Ruff, C., <b>Ugazio, G.</b>, and Fehr, E. (In Prep.) The Causal Role of the LPFC in Social Norms Compliance</li> <li>• <b>Ugazio, G.</b>, Heinzelmann, N., and Tobler, P. N. (Under Review) Practical Implications of Empirically Studying Moral Decision-Making</li> <li>• <b>Ugazio, G.</b>, Majdandzic, J., and Lamm, C. (In Prep) Are Empathy and Morality Linked? Evidence from Moral Psychology, Social and Decision Neuroscience, and Philosophy. In Malbom. H., <i>Empathy and Morality</i>, Oxford University Press</li> </ul>
<b>Invited Lectures</b>	<ul style="list-style-type: none"> <li>• <b>Moral Decisions, Emotions and Cognitive Control, University of Vienna</b></li> </ul>
<b>Media Appearances</b>	<ul style="list-style-type: none"> <li>• RSI, Swiss Italian Radio, interview on neuro-economics.</li> </ul>
<b>Conferences Participation</b>	<ul style="list-style-type: none"> <li>• Society for Neuroeconomics: Poster Presentation (Miami, September 2012)</li> <li>• OHBM 2012: Poster Presentation (Beijing, June 2012)</li> <li>• ZNZ Annual Symposium: Poster Presentation (University of Zurich, September 2011)</li> <li>• NCCR Annual Meeting: Poster Presentation (University of Geneva, July 2011)</li> <li>• HABITVS PROJECT: 'The Neuroscience of Moral Action' (Caltec, May 2011)</li> <li>• 'The economy of the Soul: Rational Choice and Moral Decision-Making (LSE, October 2008)</li> </ul>
<b>Awards</b>	<ul style="list-style-type: none"> <li>• 2012: Travel Grant, SAGW.</li> <li>• 2012: Travel Grant by the Department of Medicine, University of Zurich;</li> <li>• 2012: CanDoc ForschungKredit Grant (75.000CHF)</li> <li>• 2006: Scholarship, Carnegie Mellon University, Summer School;</li> </ul>
<b>Other Competences</b>	<ul style="list-style-type: none"> <li>• Sound Knowledge of Matlab (Mathworks) and Matlab-based toolboxes, in particular SPM8 and cogent.</li> <li>• Excellent knowledge of Office applications</li> <li>• Excellent knowledge of statistical software SPSS 17 and Stata 7</li> </ul>